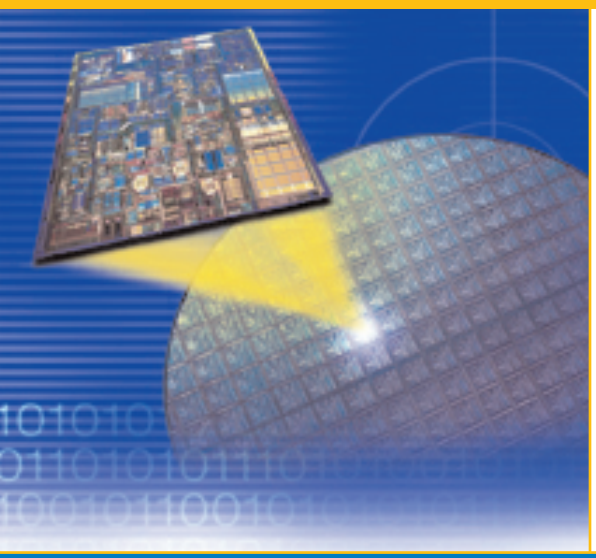
Volume 06 Issue 02 Published, May 16, 2002 ISSN 1535766X



Intel® Technology Journal

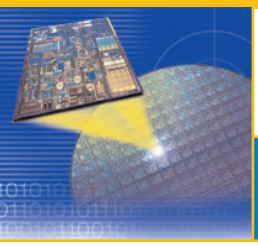
Semiconductor Technology and Manufacturing

This issue of the Intel Technology Journal describes Intel's state-of-the-art logic and flash-memory technologies and how some of the key technology elements will evolve in the near future.

Inside you'll find the following papers:

More information, including current and past issues of Intel Technology Journal, can be found at:
http://developer.intel.com/technology/itj/index.htm

Volume 06 Issue 02 Published, May 16, 2002 ISSN 1535766X



Intel® Technology Journal

Semiconductor Technology and Manufacturing

Articles

Preface

Since the invention of the integrated circuit some forty years ago, engineers and researchers around the world have worked on how to put more speed, performance and value onto smaller chips of silicon. By the end of this decade (2010) we at Intel want to reach the goal of 10 billion transistors on a single chip. This is a big challenge. Today we continue to break barriers to reach this goal. This issue (Q2, 2002, Vol. 6 Issue 2) of the *Intel Technology Journal* gives a detailed look into the exciting advances in the areas of transistor architecture, interconnects, dielectrics, lithography, and packaging.

This past year there have been many recent fundamental breakthroughs, particularly in five areas. Here we summarize some of those breakthroughs.

Transistor size: Intel's research labs have recently shown the world's smallest transistor, with a gate length of 15nm. We continue to build smaller and smaller transistors that are faster and faster. We've reduced the size from 70 nanometer to 30 nanometer to 20 nanometer, and now to 15 nanometer gates.

Manufacturing process: A new manufacturing process called 130 nanometer process technology (a nanometer is a billionth of a meter) allows Intel today to manufacture chips with circuitry so small it would take almost 1,000 of these "wires" placed side-by-side to equal the width of a human hair. This new 130-nanometer process has 60nm gate-length transistors and six layers of copper interconnect. This process is producing microprocessors today with millions of transistors and running at multi-gigahertz clock speeds.

Wafer size: Wafers, which are round polished disks made of silicon, provide the base on which chips are manufactured. Use a bigger wafer and you can reduce manufacturing costs. Intel has begun using a 300 millimeter (about 12 inches) diameter silicon wafer size, up from the previous wafer size of 200mm (about 8 inches). 300 millimeter is the size of a medium pizza in the United States, up from the previous size of a small pizza!

Lithography: Lithography is the technology used to 'print' intricate patterns that define circuits on silicon wafers. With our extreme ultraviolet (EUV) program, we've made a fundamental breakthrough in the area of lithography. EUV lithography is the technology that allows printing of lines smaller than 50nm. A few years ago, we realized that the light spectrums we were using were no longer scalable. We needed the shorter wavelengths of extreme ultraviolet beams. But rather than magnifying the beam through a glass lens as before, we now use mirrors. About five years ago we launched the industry consortium for EUV, and this year we demonstrated the first EUV using mirroring techniques.

Packaging: A silicon chip is useless without its package. The package delivers the power the chip needs and transfers all the information into and out of the chip. BBUL ("Bumpless Build-Up Layer") packaging is a new microprocessor packaging technology that has been developed by Intel. It is called bumpless because, unlike today's packages, it does not use tiny solder bumps to attach the silicon die to the package wires. Instead of having the die on top, the die is embedded in the package. It has build-up layers because the package is "grown" (built up) around the silicon die rather than being manufactured separately and bonded to it. This package is smaller, improves package inductance characteristics, and is better for multi-chip packaging.

The seven papers in this Q2, 2002, issue of Intel Technology Journal discuss the details on fundamental advancements of silicon process and manufacturing, including improvements in current technologies of 130nm logic technology, manufacturing using 300mm wafers, flash memory, digital CMOS integrated with analog RF signal elements, and next-generation advancement underway in lithography, transistor structure, and packaging technologies.

Foreword

The semiconductor industry has made phenomenal progress since Robert Noyce invented the integrated circuit over 40 years ago. The fundamental driver has been the continued shrinking of feature sizes, allowing the exponential growth in device count that tracks the well-known Moore's Law first formulated by Intel co-founder Gordon Moore. Shrinking feature sizes allow more transistors to be packed onto a piece of silicon, with each one running at higher speeds. This combination translates into more computing capabilities, ultimately delivering better value to the end user. This exponential trend has driven the amazing computing and communications revolution that is profoundly changing our world. By most measures, the industry has progressed further than anyone imagined even as recently as 10 years ago.

Making these increasingly dense and varied integrated circuits requires progress in many disciplines. New transistor materials and structures are required in order to meet new performance, speed and power objectives. New types of interconnect are required to speed signal transmission between devices. Lithography–the process of printing the intricate patterns on silicon–must break new barriers as feature sizes become ever smaller. Packaging also must become much more sophisticated to meet ever more stringent thermal management, power delivery, interconnect density and integration requirements. And all of these goals must be achieved in a cost-effective manner amenable to high-volume manufacturing.

Intel has been at the forefront of our industry since our founding in 1968, and today holds a leadership position with high-performance microprocessors, dense flash memories, and the ability to manufacture these very complex products in high volume. This issue of the Intel Technology Journal describes Intel's state-of-the-art logic and flash-memory technologies and how some of the key technology elements will evolve in the near future.

130nm Logic Technology Featuring 60nm Transistors, Low-K Dielectrics, and Cu Interconnects

Scott Thompson, Technology and Manufacturing Group, Intel Corporation
Mohsen Alavi, Technology and Manufacturing Group, Intel Corporation
Makarem Hussein, Technology and Manufacturing Group, Intel Corporation
Pauline Jacob, Technology and Manufacturing Group, Intel Corporation
Chris Kenyon, Technology and Manufacturing Group, Intel Corporation
Peter Moon, Technology and Manufacturing Group, Intel Corporation
Matthew Prince, Technology and Manufacturing Group, Intel Corporation
Sam Sivakumar, Technology and Manufacturing Group, Intel Corporation
Sunit Tyagi, Technology and Manufacturing Group, Intel Corporation
Mark Bohr, Technology and Manufacturing Group, Intel Corporation

Index words: CMOS transistor, logic technology, copper interconnects

ABSTRACT

Transistor gate dimensions have been reduced 200X during the past 30 years (from 10 m in the 1970s to a present-day size of 0.06 m). The transistor and feature size scaling have enabled microprocessor performance to increase exponentially with transistor density and microprocessor clock frequency doubling every two years. In this paper we describe Intel's latest 130nm CMOS logic technology used to make high-performance microprocessors >3GHz.

INTRODUCTION

For more than 30 years, MOS device technologies have been improving at a dramatic rate [1-6]. A large part of the success of the MOS transistor is due to the fact that it can be scaled to increasingly smaller dimensions, which results in higher performance. The ability to consistently improve performance while decreasing power consumption has made CMOS architecture the dominant technology for integrated circuits. The scaling of the CMOS transistor has been the primary factor driving improvements in microprocessor performance. Transistor delay times have decreased by more than 30% per technology generation resulting in a doubling of microprocessor performance every two years. Recently, chip performance has also come to be limited by back-end RC delay if low-resistance metal lines or low dielectric constant interlayer dielectrics are not used.

In this paper we describe Intel's 130nm logic technology that features 60nm gate length and 1.5nm gate-oxide transistors for high-performance and low-k interdielectrics with six layers of Cu interconnects. We first discuss transistor scaling. Next, we present data from our 130nm technology on 60nm transistors and copper interconnects with low-k Fluorinated SiO_2. We conclude with static random access memory (SRAM) and microprocessor performance data.

TRANSISTOR SCALING OVERVIEW

Transistor scaling has been the key driving force behind the rapid increase in microprocessor clock frequency. Figure 1 shows the scaling trend of clock frequency. The technology target for the 130nm node was to produce microprocessors at >3GHz.

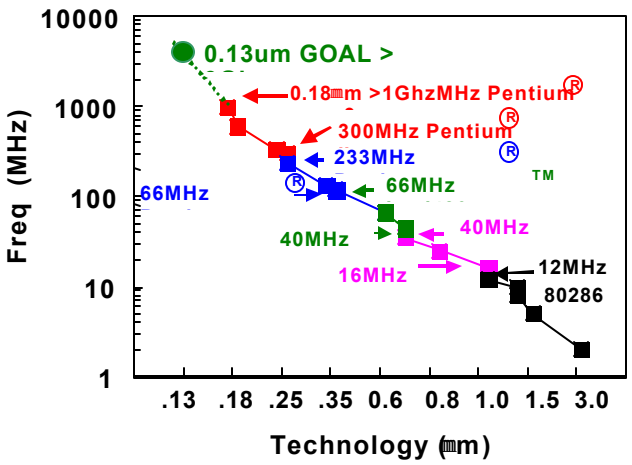


Figure 1: Microprocessor clock frequency vs. technology generation

Two primary factors set the transistor speed and hence microprocessor clock frequency for a given design: transistor channel length and gate-oxide thickness (Figure 2). To reach the >3GHz goal, circuit simulations show that 60nm gate length and 1.5nm gate-oxide thickness are required for the 130nm technology node. The 60nm transistor requires a significant acceleration of the transistor feature size relative to the technology and light source.

Figure 3 shows the trends of these key feature sizes versus technology generation. The 130nm technology node was designed for the fabrication of Intel Pentium® 4 microprocessors in high-volume manufacturing. Once the Pentium 4 chip architecture is set, the transistor speed required for 3GHz operation can be determined. To obtain a clock frequency of >3GHz, it was determined that a 1.3mA/ m transistor saturation drive current would be needed. This value of drive current is significantly higher than the value in our 180nm technology (~1.0mA / m). 60nm transistors with 1.5nm physical oxide thickness will allow for CV/I close to 1ps (Figure 4) and saturation drive current of 1.3mA/ m (Figure 6). This drive current is the highest to date in high-volume production. Key to obtaining the high drive current is high channel mobility. The channel mobility decreases at higher effective oxide fields for the smaller feature size technology. The electron mobility is shown in Figure 5. The electron mobility is on the universal mobility curve even though the physical thickness of the oxide is only 1.5nm.

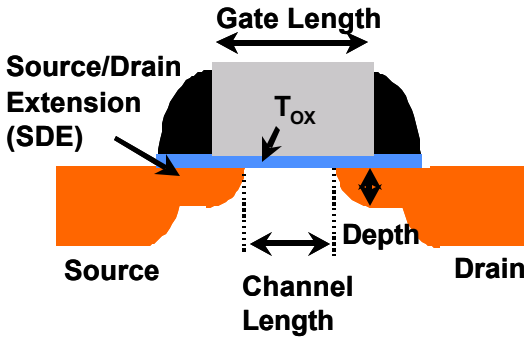


Figure 2: Cross-section drawing of a CMOS transistor

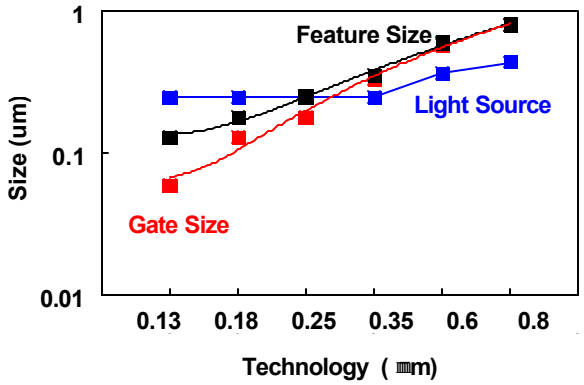


Figure 3: Technology feature size, wavelength light source, and transistor gate size vs. technology node

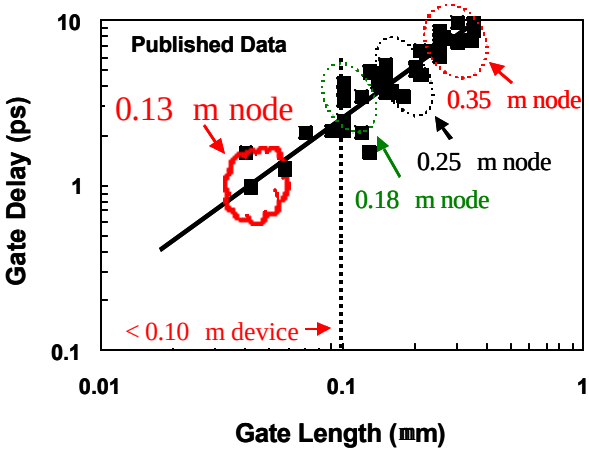


Figure 4: CV/I gate delay vs. transistor gate length

Intel and Pentium are registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

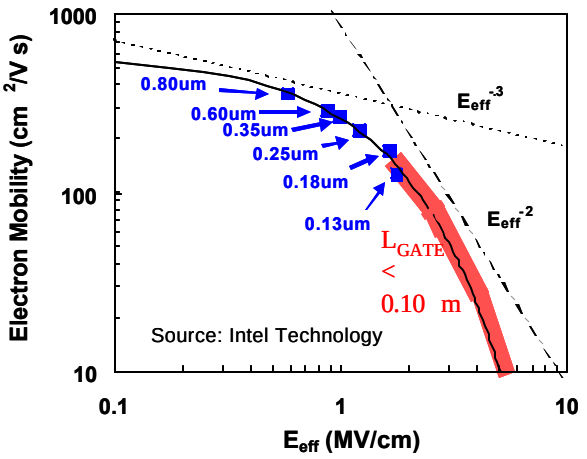


Figure 5: Electron mobility vs. effective vertical electrical field

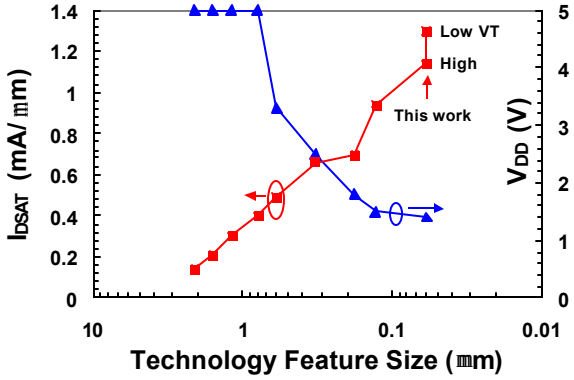


Figure 6: Transistor saturated drive current vs. technology feature size

Process Flow and Technology Features

Front-end technology features include shallow trench isolation, retrograde wells, shallow abrupt source/drain extensions, halo implants, deep source/drain, and cobalt salicidation. Figure 7 shows a front-end cross section of the technology. The minimum pitches and thicknesses for the technology layers are summarized in Table 1. The rules enable a 2.0 um ^{2} 6-T SRAM cell (1.22 x 1.64 um). Figure 8 shows a top-down scanning electron micrograph (SEM) of the polysilicon gate conductor and the Metal 1 connections. The interconnect technology uses dual damascene copper to reduce the resistances of the six layers of interconnects. Fluorinated SiO_2 is used as an inter-level dielectric (k is measured to be 3.6).

Table 1: Layer pitch, thickness (nm) and aspect ratio

LAYER	PITCH	THICK	AR
Isolation	345	450	-
Polysilicon	319	160	-
Metal 1	293	280	1.7
Metal 2,3	425	360	1.7
Metal 4	718	570	1.6
Metal 5	1064	900	1.7
Metal 6	1143	1200	2.1

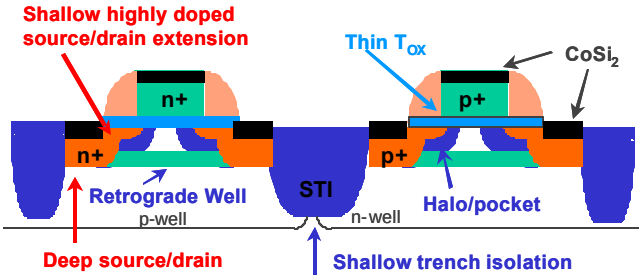


Figure 7: Cross-section drawing of 130nm technology front-end

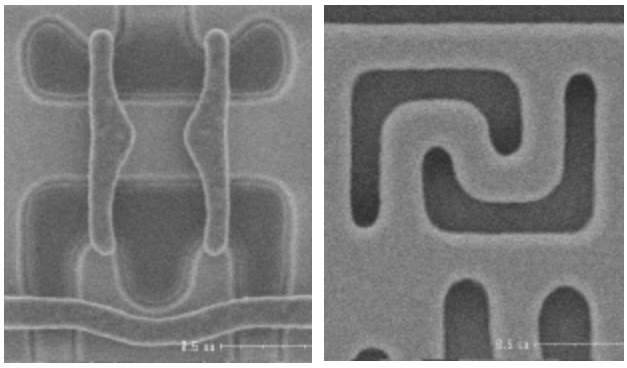


Figure 8: Top-down SEM of polysilicon gate conductor and Metal 1 connections

TRANSISTOR FEATURES

(a) Gate Length Dimension: Figure 9 shows a cross-sectional transmission electron micrograph (TEM) for a transistor with a 60nm gate length and straight poly-Si sidewall profile as opposed to the notched poly used in the 180nm node [6]. Straight sidewall gates were chosen at the 130nm node since the source drain extension does not have to diffuse under the notch, thus allowing for shallower junctions to be fabricated.

At aggressive gate lengths of 60nm, controlling short channel effects at low-threshold voltage, by using shallow junctions and abrupt halo doping, is key to achieving high linear and saturation drive currents.

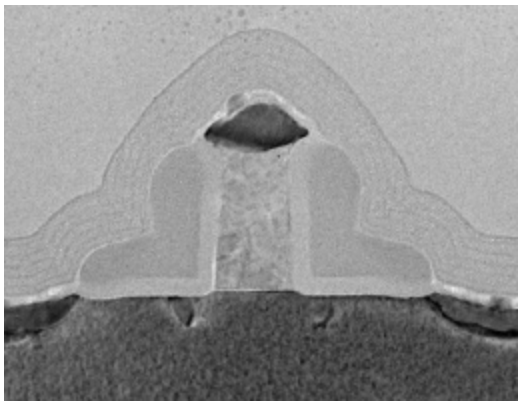


Figure 9: TEM cross section of 60nm NMOS

(b) 1.5nm Physical Gate-Oxide: In order to achieve high drive current and minimize short channel effects, a gate-oxide process with a 1.5nm physical thickness was developed that meets performance, reliability, and manufacturability criteria (Figure 10). High-electron and hole mobilities are required to achieve high linear drive current, which can be missed in technology optimization, since transistor linear current is not reflected in a simple CV/I metric. Concerns have been raised that in ultra-thin oxides, gate-electrode-to-oxide interface scattering and high fixed charge due to nitridation reduce mobility. The measured mobility dependence on the effective oxide field, shown in Figure 11, demonstrates that high-electron and hole mobilities can be achieved for well-optimized gate oxides with a thickness of 1.5nm.

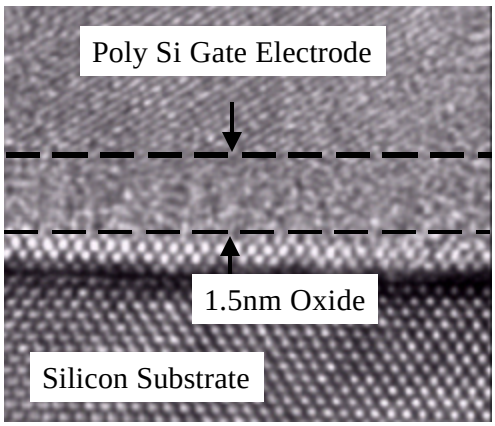


Figure 10: TEM of 1.5nm physical gate oxide

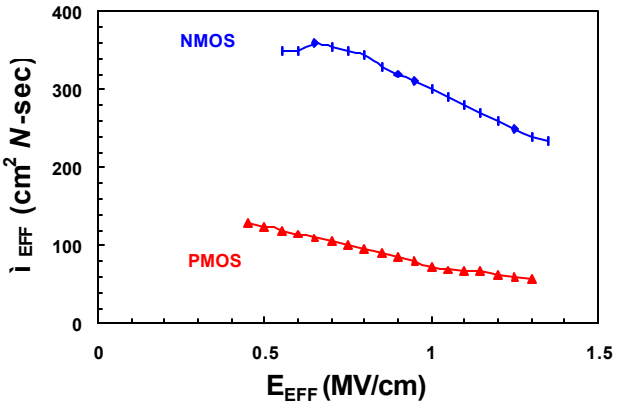


Figure 11: Mobility vs. effective electric field

(c) Well-Halo and SD-Extension Engineering: A simple but ineffective way to offer high-saturation drive current at small gate lengths is to use high well-doping to raise the threshold voltage to control short channel effects. This approach offers low CV/I but does not improve product performance, for two reasons. First, the linear drive current will be significantly degraded (saturated drive current is not degraded at a fixed I_{OFF} due to high drain-induced barrier lowering (DIBL)). Second, the high well-doping leads to increased threshold voltage variations due to gate length variation (present in the range of +/- 10% L_{GATE} for a modern technology). In this work we use retrograde wells, and low-energy, high-angle abrupt halo implants with shallow junctions formed by low-diffusion processing to control short channel effects. Figure 12 shows the N-channel threshold voltage versus gate length resulting in a linear threshold voltage of 300 and 270mV at a gate length of 60nm for the high- and low-threshold devices, respectively. From Figure 12, DIBL for the 60nm NMOS devices is measured to be <100mV/V for high- and bw-threshold devices. Similar results have been achieved for p-channel devices.

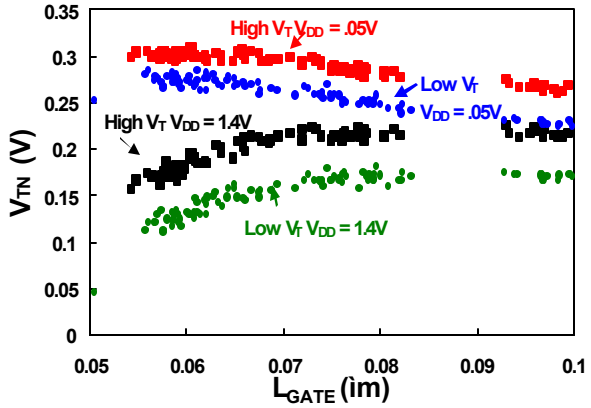


Figure 12: V_{TN} vs. L_{GATE}

High V_T saturation drive currents are 1.14mA/um for N-channel and 0.56mA/um for P-channel devices (Figure 13). Low V_T drive currents are 1.30mA/um for N-channel and 0.66mA/um for P-channel devices (Figure 14). Sub-threshold slopes for both N-channel and P-channel high- and low-threshold devices remain well controlled at less than 85mV/decade at L_{GATE}=60 nm (Figure 15). The I_{ON}/I_{OFF} ratio remains high for the aggressively scaled power supply voltage of 1.4V (Figure 16). Table 2 shows the transistor I_{ON} and I_{OFF} at 0.7 and 1.4 V.

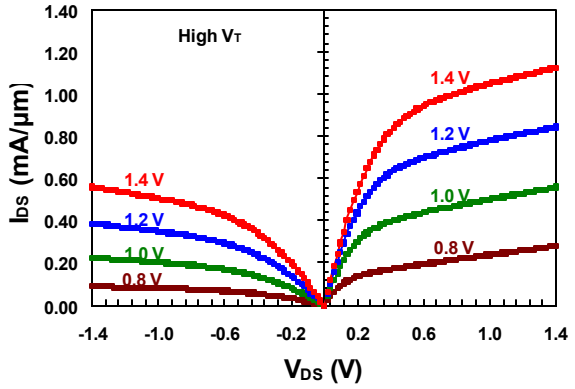


Figure 13: I-V curves for high V_T device (L_{GATE}=60nm)

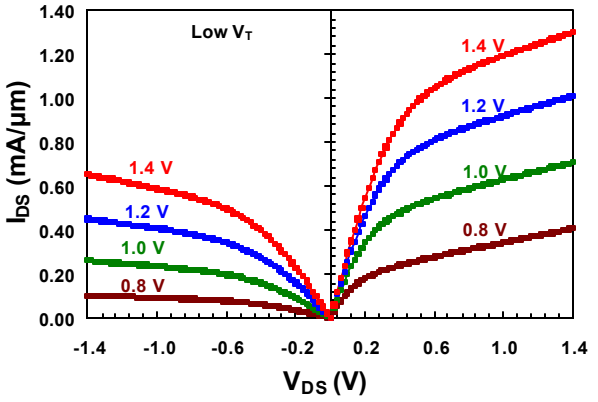


Figure 14: I-V curves for low V_T device (L_{GATE}=60nm)

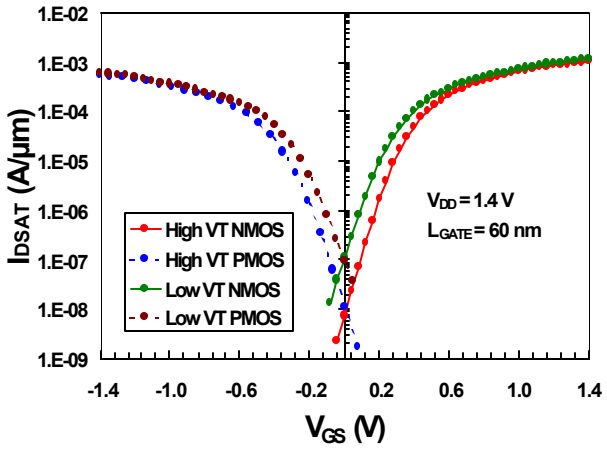


Figure 15: Sub-threshold characteristics

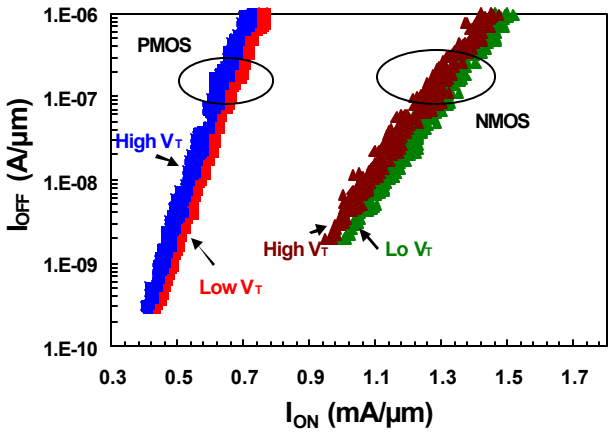


Figure 16: I_{ON} Vs I_{OFF} (V_{DD}=1.4V)

Table 2: I_{ON} and I_{OFF} at 0.7 and 1.4V V_{DD}

DEVICE	VDD (V)	I_{OFF} (N) (nA/um)	I_{ON} (N) (mA/um)	I_{ON}(P) (mA/um)
Low V_T	1.4	100	1.30	0.66
High V_T	1.4	10	1.14	0.56
Low V_T	0.7	20	0.37	0.19
High V_T	0.7	2	0.32	0.16

In a modern microprocessor with six layers of interconnects, transistor loads are comprised of >50% interconnect capacitance. To obtain high product performance it is necessary to provide transistors with more than low CV/I; you also need high saturation and linear drive currents. Figure 6 shows the recent trend of saturation drive currents for Intel's process technologies. This work extends the trend to offer the highest drive current to date of 1.30mA/um for low-threshold N-channel devices.

INTERCONNECTS

Chip performance is increasingly limited by the RC delay of the interconnect as the transistor delay progressively decreases while the narrower lines and space actually increase the delay associated with interconnects. Using copper interconnects helps reduce this effect. This process technology uses dual damascene copper to reduce the resistances of the interconnects. Fluorinated SiO_2 (FSG) is used as an inter-level dielectric (ILD) to reduce the dielectric constant; the dielectric constant k is measured to be 3.6. Figure 17 is a cross-section Scanning Electron Micrograph (SEM) image showing the dual damascene interconnects.

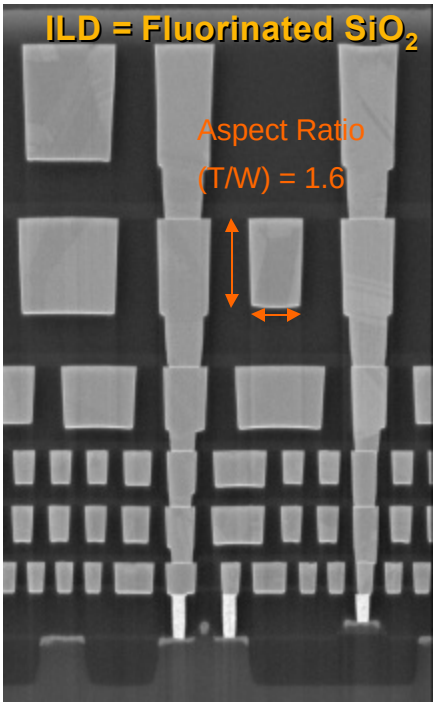


Figure 17: Cross-section SEM image of a processed wafer

Table 1 lists the metal pitches. The pitch is 350nm at the first metal layer and increases to 1200nm at the top layer. Metal aspect ratios are optimized for minimum RC delay and range from 1.6 to 2. The first metal layer uses a single damascene process, and tungsten plugs are used as contacts to the silicided regions on the silicon and poly-silicon. Unlanded contacts are supported by using an Si_3N_4 layer for a contact etch stop. Copper interconnects are used because of the material's lower resistivity. The advantage is seen in Figure 18, where the sheet resistance is shown as a function of the minimum pitch of each metal layer and compared to earlier results from 180nm technologies using Al [6] and Cu [6]. The present technology exhibits 30% lower sheet resistance at the same metal pitch due to the use of Cu with high aspect ratios. The total line capacitance is 230fF/mm for M1 to M5 and slightly higher for the top layer.

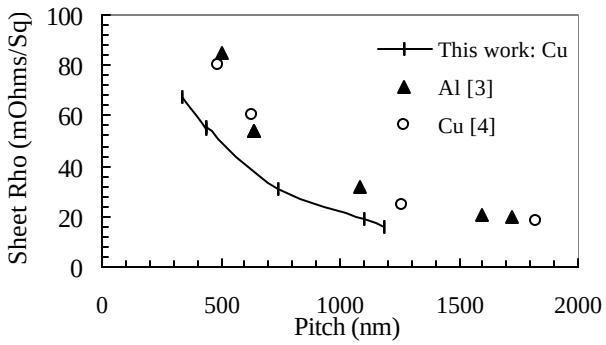


Figure 18: Sheet resistance as a function of layer pitch

To benchmark the performance of interconnects, Figure 19 shows the RC delay in picoseconds per millimeter of wire. Data for each metal layer are shown as a function of the minimum pitch at that layer. For a given pitch, 50% reduction in RC is achieved by using Cu interconnects and FSG ILD.

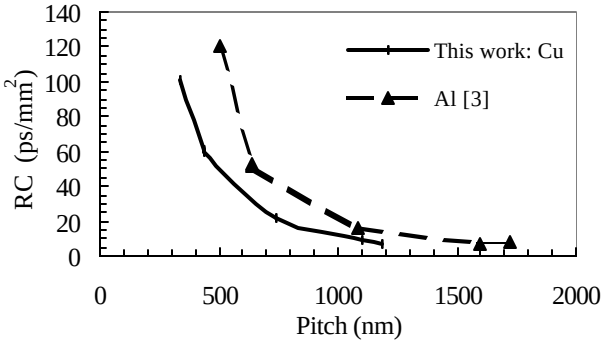


Figure 19: RC delay for a wire length of 1mm as a function of layer pitch

Performance Metrics

Figure 20 shows measured inverter gate delay versus n-channel off-state leakage for an unloaded ring oscillator (fan out =1) operating at 1.4V at room temperature. PMOS off-state leakage is fixed at 10nA/um for these devices. The delay per stage at 1.4V falls below 6psec when the off-state leakage is about 10nA/um.

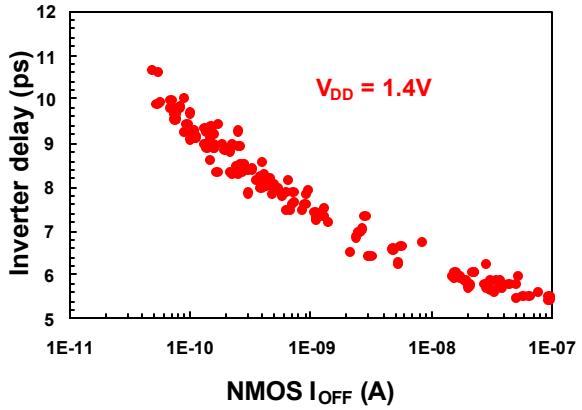


Figure 20: Inverter delay (PMOS I_{OFF} = 10nA/um)

Power consumption is a growing concern for high-performance microprocessors with increasing clock frequency and transistor count. The best way to reduce power is to operate at a low supply voltage. Figure 21 shows that by improving device matching and eliminating defects that cause device mismatches, an 18Mb SRAM fabricated in this technology can operate at voltages of down to 0.5V. A metric, which comprehends both power and speed, is the energy-delay product. Figure 22 shows the estimated NMOS energy-delay product for a large number of published devices and for the devices reported in this paper. As evident from Figure 22, the NMOS energy-delay product is better than the published industry trend.

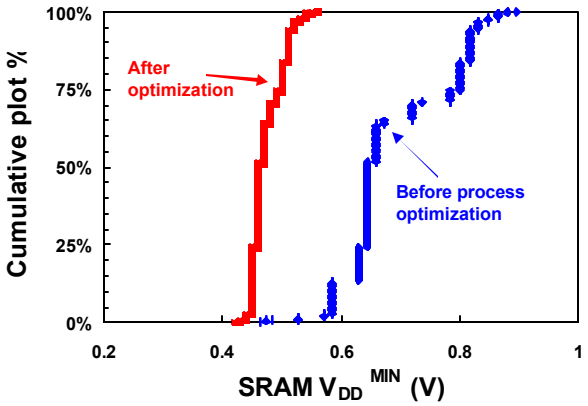


Figure 21: SRAM operation vs. voltage

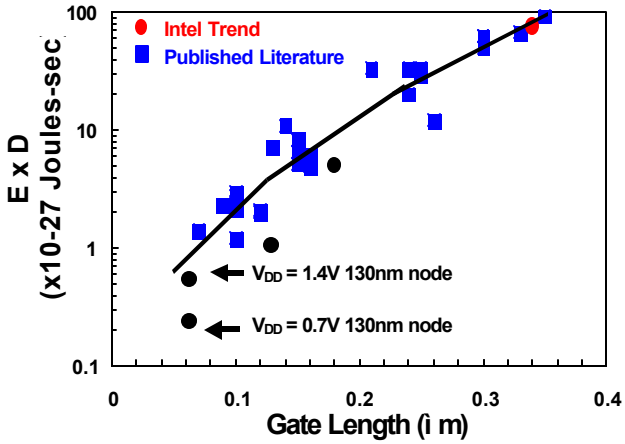


Figure 22: Energy-delay product vs. L_{GATE}

An 18 Mbit CMOS SRAM, Pentium III and Pentium® 4 microprocessor were fabricated and used as yield and reliability test vehicles during the process development. Figure 23 shows the die photo of the Pentium 4 in the 0.18 and 0.13um technologies. The SRAM and microprocessor die yields are equivalent or better than past technologies at this point of time relative to ramping in high-volume manufacturing. The performance of the Pentium 4 processor is measured using the maximum clock frequency of operation. Figure 24 shows the schmoo plot for the Pentium 4, i.e., the maximum frequency as a function of voltage. At an operation voltage of 1.4V, the present design version of the Pentium 4 microprocessor has a clock frequency of 2.5GHz.

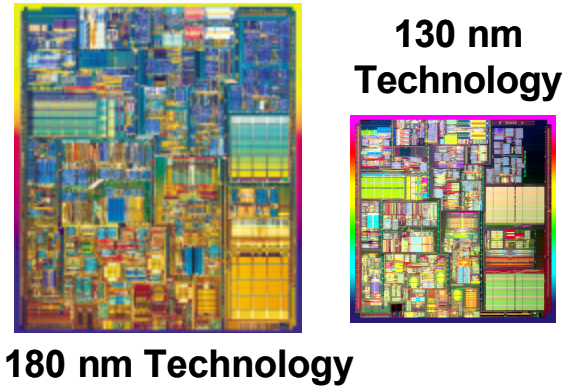


Figure 23: Comparison of 180nm technology to 130nm technology

Pentium is a registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

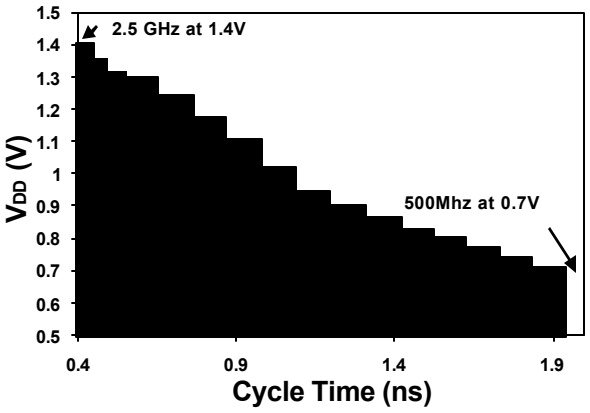


Figure 24: Fmax schmoo plot for the Pentium® 4 processor

CONCLUSIONS

A 130nm-generation logic technology has been developed and is in high-volume manufacturing with high-performance transistors that can operate in the range of 0.7 and 1.4 V. The technology performance capabilities are demonstrated with ring oscillator delays of 6 ps/stage and with a Pentium 4 processor operating at 2.5 GHz. The transistors can support microprocessors operating at >3GHz.

ACKNOWLEDGMENTS

The authors acknowledge the collaborative efforts of our colleagues in the Portland Technology Development Group, the Technology Computer Aided Design Group, and in the Corporate Quality and Reliability group.

REFERENCES

[1] M. Bohr, S.U. Ahmed, L. Brigham, R. Chau, R. Gasser, R. Green, W. Hargrove, E. Lee, R. Natter, S. Thompson, K. Weldon and S. Yang, *IEDM Technical Digest*, 1994, p. 273.

[2] M. Bohr, S.S. Ahmed, S.U. Ahmed, M. Bost, T. Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S. Sivakumar, S. Thompson, J. Tsai, and S. Yang, *IEDM Technical Digest*, 1996, p. 847.

[3] S. Thompson, VLSI Symposium Technology Short Course, 1998.

[4] S. Thompson, P.A. Packan, and M.T. Bohr, *VLSI Symposium Digest*, 1996, p. 154.

[5] C.T. Sah, Fundamentals of Solid-State Electronics, 1991, p. 553.

[6] S. Thompson, M. Alavi, R. Arghavani, A. Brand, R. Bigwood, J. Brandenburg, B. Crew, V. Dubin, M. Hussein, P. Jacob, C. Kenyon, E. Lee, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, M. Prince, R. Schweinfurth, S. Sivakumar, P. Smith, M. Stettler, S.Tyagi, M. Wei, J. Xu, S. Yang and M. Bohr, *IEDM Technical Digest*, 2001, p. 11.6.1-11.6.4.

AUTHORS' BIOGRAPHIES

Scott Thompson joined Intel in 1992 after completing his Ph.D., under Professor C. T. Sah at the University of Florida, on thin gate oxides. He has worked on transistor design and front-end process integration on Intel's 0.35, 0.25, 0.18, and 0.13 m silicon process technology design for the Intel® Pentium® and the Pentium® II microprocessors. Scott is currently managing the development of Intel's 90nm logic technology. His e-mail is scott.thompson@intel.com.

Mohsen Alavi joined Intel in 1986 after completing his Ph.D. in Electrical Engineering at Michigan State University on Schottky Barrier Diodes. He has worked on transistor and interconnect development and reliability of many of Intel's logic process technologies starting from the 1um process. More recently, he has been the reliability program manager for 0.13um and subsequently, 90nm logic technology development and is currently manager of LTD Q&R. His e-mail is mohsen.alavi@intel.com.

Makarem Hussein is a Principal Engineer with Patterning Area of Portland Technology Development. He graduated from the University of Wisconsin-Madison in 1990 with a Ph.D. degree in Nuclear Engineering and Engineering Physics. He joined Intel in 1992, and since then has been working on developing dry etch processes. His most recent focus has been on the patterning of dielectric substrates for copper interconnect systems. He holds six US patents and has authored/co-authored more than 15 articles in the field of plasma etching and patterning technology. His e-mail is makarem.hussein@intel.com

Pauline Jacob joined Intel in 1994 after completing a Ph.D. in Chemical Engineering at the University of Washington. She has worked on diffusion process development since Intel's 0.35um process technology. Pauline is currently the diffusion group leader working on the development of Intel's 90nm gate-oxide module. Her e-mail is pauline.n.jacob@intel.com.

Chris Kenyon is a lithography group leader in Intel's Logic Technology Development organization. He joined Intel in 1996 and has worked primarily on Intel's gate patterning process since that time. He is currently responsible for developing the gate patterning process for the 90nm CMOS node. He received his B.A. degree from Princeton University in 1990 and his Ph.D. degree from Caltech in 1996 in Physical Chemistry. His e-mail is Chris.Kenyon@intel.com

Peter Moon joined Intel in 1988 after completing his Ph.D. in Materials Science at the Massachusetts Institute of Technology. He has worked on process integration for Intel's 0.8, 0.35 and 0.13um silicon process technologies for Pentium® microprocessors including Intel's first use of shallow trench isolation (0.35um) and Intel's first use of copper interconnects (0.13um). Peter is currently leading the development of Intel's interconnect process for the 45nm process generation. His e-mail is Peter.Moon@intel.com

Sam Sivakumar joined Intel in 1990 after graduating from the University of Illinois. He is a member of the Portland Technology Development lithography group and has worked on patterning process development for a variety of Intel's logic processes. He is currently responsible for lithography development for Intel's 90nm logic process. His e-mail is sam.sivakumar@intel.com.

Mark T. Bohr joined Intel in 1978 after receiving an M.S.E.E. degree from the University of Illinois. He has been a member of the Portland Technology Development group since 1978 and has been responsible for process integration and device design on a variety of DRAM, SRAM, and logic technologies, including recent 0.35um and 0.25um logic technologies. He is an Intel Fellow and director of process architecture and integration. He is currently directing development activities on 0.18um and 0.13um logic technologies. His e-mail is mark.bohr@intel.com.

Matthew Prince joined Intel in 1988 after graduating from Clarkson University in New York. Since 1989 he has developed ILD, W, and Cu CMP technologies. His e-mail address is matthew.j.prince@intel.com.

Copyright © Intel Corporation 2002. This publication was downloaded from http://developer.intel.com/ .

Legal notices at
http://www.intel.com/sites/corporate/tradmarx.htm

Process Development and Manufacturing of High-Performance Microprocessors on 300mm Wafers

Sanjay Natarajan, Logic Technology Development, Intel Corporation
Melton Bost, Logic Technology Development, Intel Corporation
Derek Fisher, Logic Technology Development, Intel Corporation
David Krick, Logic Technology Development, Intel Corporation
Chris Kenyon, Logic Technology Development, Intel Corporation
Chris Kardas, Logic Technology Development, Intel Corporation
Chris Parker, Logic Technology Development, Intel Corporation
Robert Gasser, Jr., Logic Technology Development, Intel Corporation

Index words: 300mm, Px60, P1260, 0.13 m, 130nm, Copy Exactly!

ABSTRACT

Over 35 years ago, Moore's Law established the nature of competition in the semiconductor industry by projecting a 2x transistor density improvement approximately every 18 months. Faced with increasingly challenging process technology issues, industry leaders such as Intel have had to achieve increasingly faster yield improvement and volume production ramps to maintain competitiveness. The Copy Exactly! methodology, which has been used since 1992 to transfer technologies and ramp new factories, has been instrumental in allowing Intel to meet these challenges.

The subject of this paper is the successful extension of Copy Exactly! to Intel's first 300mm process technology, P1260, to achieve rapid yield learning and volume production. P1260 replicates Intel's industry-leading 200mm 0.13 m CMOS process in performance, yield, reliability, and density, with SRAM cell sizes below 2 m^2 [1]. Intel has used the Copy Exactly! methodology for several generations with documented success, and we present perhaps the most compelling evidence to date of its utility: accurate replication of an industry-leading 200mm 0.13 m CMOS process on a 300mm wafer size using a completely new process equipment set.

INTRODUCTION

Moore's Law

In 1965, Gordon Moore, then R&D manager at Fairchild Semiconductor and now Chairman Emeritus of Intel Corporation, characterized the rate of progress in the semiconductor industry and arrived at an astounding conclusion: the density of transistors per integrated circuit (IC) had been doubling at regular intervals and would continue to do so indefinitely [2].

The observation, later termed "Moore's Law," has been extremely influential in the semiconductor industry, even to the point of becoming self-fulfilling. Since Moore's Law has accurately predicted past IC growth, it is also viewed as a method for predicting future trends, setting goals for innovation, directing the pace of the technology treadmill, and ultimately defining the nature of industry competition [3].

Delivering the regular progress dictated by Moore's Law in the face of increasingly complex process technologies requires steady improvements in the pace of yield learning and volume manufacturing capability. Figures 1 and 2 illustrate this trend for Intel's process technologies. Figure 1 shows the steadily increasing rate of production ramp for each of the last six process generations. Across these six generations, there has been a 4x increase in the ramp rate, measured in wafer starts per week per Fab. In addition, this increase has been achieved across more Fabs each generation. The net result is a greater than 20x

increase in normalized die output in early ramp over the past six generations. Figure 2 illustrates the rapid increase in yield-learning trends over the last seven generations. The graph shows defect learning rates (the y-axis is the logarithm of defect density, so lower is better) for Intel technologies from the start of process development through initial production. There are three key points in this data. First, the elapsed time from the start of development to the point of high yield is decreasing with subsequent technology generations. Second, the inflection point, where yield learning slows down, is occurring at higher yields with subsequent generations. Finally, the time between new process introductions is decreasing. The net result is a greater than 5x increase in normalized good die per wafer at the start of production, over the past seven generations.

These continuously increasing ramp rates and ever-improving yield-learning rates have been instrumental in maintaining Intel's leadership in the technology race, as defined by Moore's Law. There are three primary methods that enable rapid yield learning and manufacturing ramp. The first is predictive in-line metrology to shorten the cycle time for yield improvement feedback. The second is designing the process for manufacturability and performance, including using advanced process control and developing new materials. The final method is the Copy Exactly! process for transfer and ramp. The first two methods are discussed in detail elsewhere [4]. This paper focuses on Copy Exactly!.

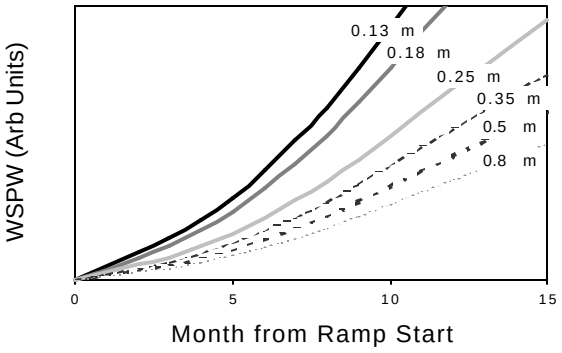


Figure 1: Intel high-volume production ramp rates

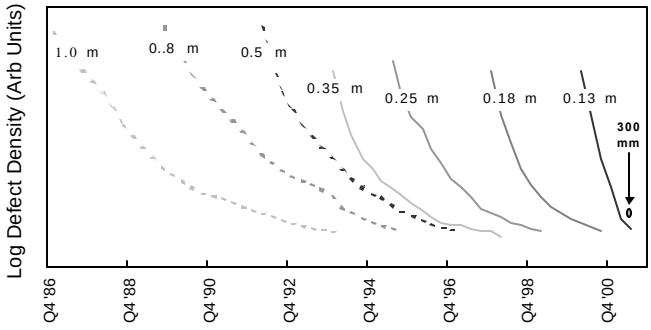


Figure 2: Intel defect density trends

COPY EXACTLY!

Up to Intel's 1 m process technology, die yields were becoming increasingly harder to match as processes were transferred from development to manufacturing facilities. During the 1 m process transfer, the first production Fab attempted to copy the development Fab closely while the second and third Fabs instituted changes (intended to be process improvements) during transfer. The results, shown in Figure 3, are striking. The so-called improvements actually resulted in an up to 10x *reduction* in die per wafer compared to the development Fab and first production Fab. This phenomenon led to the development of the Copy Exactly! methodology.

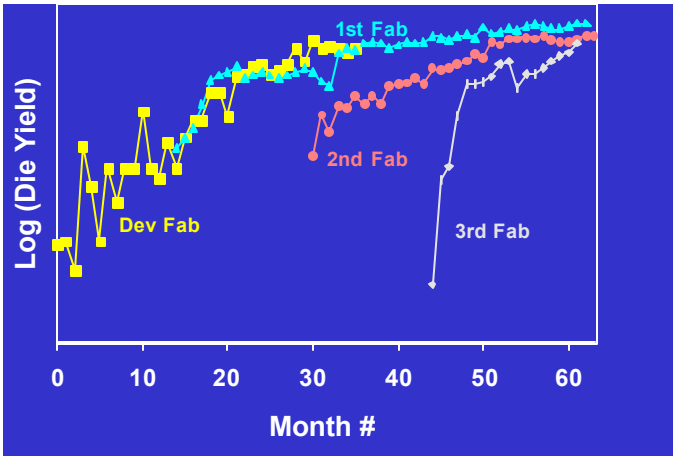


Figure 3: The birth of Copy Exactly!

The current Copy Exactly! methodology used at Intel is shown in Figure 4. The key principle behind Copy Exactly! is that Fabs running a given process technology strive to be matched in every respect except where prohibited by hard barriers. Physical inputs, such as chemical sources and purities, facilities, and hookups are all derived from the same specifications. Likewise,

equipment configurations and process recipes are matched exactly, and monitors that predict yield, reliability, and performance are all matched to within 1.5 . Once matched, changes are coordinated through cross-Fab joint engineering teams. Audits of equipment configurations and process monitors are routinely done to ensure ongoing matching. High-level tactical and strategic changes are executed in all Fabs under joint engineering and management structures.

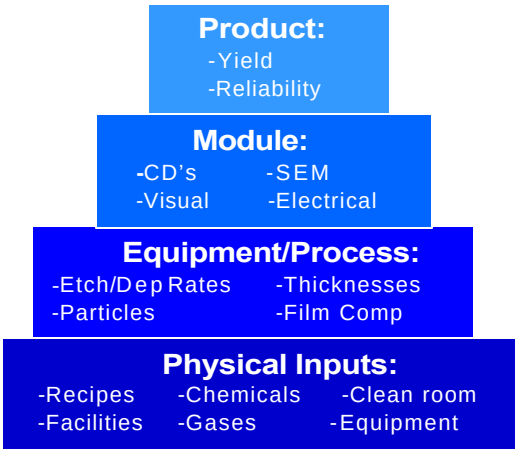


Figure 4: Current Copy Exactly! methodology

Figure 5 shows the benefit this methodology has brought since the 0.5 m technology generation. In contrast to the range of die yields observed in the 1 m generation without Copy Exactly!, every generation from the 0.5 m generation to the most recent 0.13 m generation has seen multiple Fabs started with matched die yields.

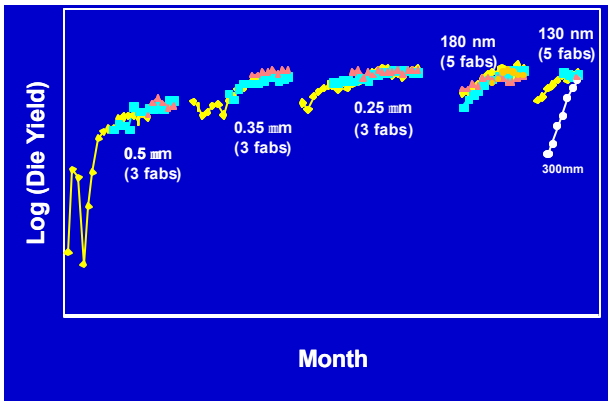


Figure 5: Die yield matching with Copy Exactly!

OVERVIEW OF INTEL'S 0.13mM LOGIC TECHNOLOGY

Most recently, Intel led the industry in 2001 with the volume manufacturing ramp of a 0.13 m CMOS technology featuring 70nm dual V_t transistors, copper and low k (dielectric constant) interconnects and 2 m^2 SRAM cell sizes [1]. Table 1 summarizes the design rules for this process technology. Figures 6 and 7 illustrate Pentium III processor die size and show the relative performance between this technology and the previous 0.18 m process generation. The transition from 0.18 m to 0.13 m process technology yields a greater than 40% increase in product frequency.

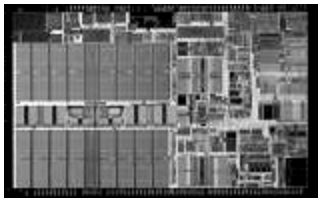


Figure 6: Pentium® III die on 0.13mm process

Layer	Pitch (nm)	Thickness(nm)	Aspect Ratio
Isolation		345 450	-
Polysilicon	319	160	-
Metal 1	293	280	1.7
Metal 2, 3	425	360	1.7
Metal 4	718	570	1.6
Metal 5	1064	900	1.7
Metal 6	1143	1200	2.1

Table 1: Intel's 0.13mm CMOS design rules

Pentium is a registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

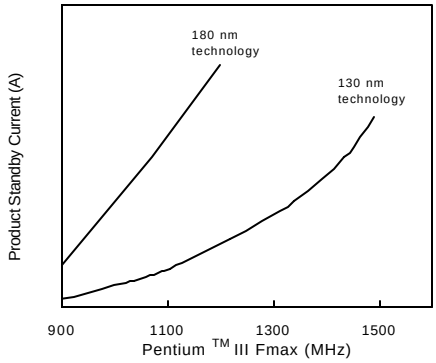


Figure 7: Pentium® III performance on 0.13mm and 0.18mm processes

300mm Wafer-Size Conversion

Intel chose the 0.13 m generation to make the wafer size change from 200mm (8") to 300mm (12"). This wafer size increase is part of an ongoing evolution beginning over 30 years ago with 1" wafers. The key driver for wafer size increase is cost reduction. The larger wafers provide a 2.25x increase in area and, due to the rectangular die size, an even larger increase in die per wafer. Manufacturing costs per wafer scale at less than this rate, so there is an overall reduction in cost per die at the larger wafer size.

The 300mm wafer size also brought a unique challenge. For the first time, the wafer size had grown large enough to pose an ergonomic hazard. A full lot of 300mm wafers weighs 18 lbs., and manual handling of 300mm wafers is prohibited due to ergonomic risks. In contrast, a full lot of 200mm wafers weighs 8 lbs. and is much smaller than a lot of 300mm wafers. 200mm wafer lots are routinely handled manually. The requirement for automated and mechanically-assisted wafer handling posed by the 300mm wafer size translates into longer cycle times for routine Fab tasks and ultimately translated into overall delays during process development.

The principal issue, however, in wafer size conversions is that the equipment set and process recipes must be completely changed to support the larger wafer. 300mm process equipment was selected using a rigorous and data-based approach. Similarity to the existing 200mm toolset was not a major factor during equipment selection: technical capability, cost, extendibility to future technologies, and productivity were. This selection process delivered a highly capable and productive toolset that could be reused for future technologies, but it drove changes away from well-characterized but less productive toolsets that had been operating, in some cases, for many years in Intel Fabs. A state-of-the-art process such as Intel's 0.13 m process has several-hundred process steps using 50-100 unique process tools. For every step, recipes must be rewritten to accommodate the larger wafers, but the higher-level goal is that the 300mm process must be essentially identical to the 200mm process in performance, reliability, and yield. With a completely new toolset and recipes that could not be simply copied or scaled, Intel faced a huge challenge in matching outputs between its 200mm and 300mm technologies. To meet the challenge, the Copy Exactly! process was adapted. This adaptation is described in the next section.

300MM COPY EXACTLY!

The development of the 300mm 0.13 m process used a modified Copy Exactly! process. Because the equipment was, by definition, different, and facility changes had to be made to accommodate the new equipment and new wafers, many of the physical inputs could not be matched. Figure 8 illustrates this. At the physical input level, recipes, facilities, equipment, and cleanroom were all not matched to 200mm. At the equipment and process level, many characteristics could not be matched because the tools either operated in different regimes from their 200mm equivalents or were based on different operating principles altogether.

However, to achieve matched output at the highest level, matching to 200mm was very extensive in other areas. To a large extent, chemicals and gases were matched, in some cases sharing a common distribution system with 200mm. Recipes were optimized for 300mm based on scaling 200mm recipes wherever possible, matching tool-level outputs to 200mm wherever possible, and always matching critical inputs to tools. "Critical inputs" are defined as those that have an impact on the wafer beyond their intended process step. For example, temperature in a thermal oxidation operation is considered a critical input because, in addition to modulating the film properties (the intended process step), temperature may also have an unintended impact on dopant diffusion and activation.

Critical outputs such as film thickness, profiles, and electrical properties were matched to 200mm within 1.5 . Variability was targeted to be equivalent or better than 200mm. The results of applying this methodology are presented in the next section.

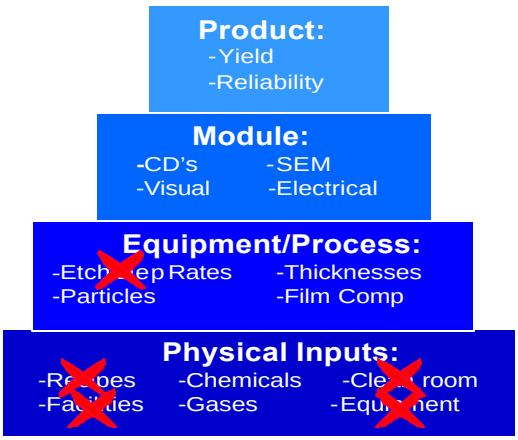


Figure 8: 200mm-to-300mm Copy Exactly! methodology

RESULTS

We now review several key metrics of our 300mm 0.13 m process and compare them to the 200mm process. We begin with module-level data, characterizing the matching of specific tools or subsets of the overall process. We then report matching data on transistor and Pentium 4 processor product performance, yield, and reliability. The data shown are a representative sample of all such indicators. In general, all data are matched between 200mm and 300mm to a similar degree. Across the board, the data show excellent matching between the 200mm and 300mm 0.13 m processes.

Module-Level Matching

Figure 9 shows within-wafer matching for a representative in-line key monitor. Shown here are 200mm and 300mm wafer maps of gate-oxide thickness. The data show that 300mm wafers have slightly better within-wafer gate-oxide thickness variation than 200mm wafers.

Figure 10 shows cumulative distributions for back-end Via resistances for 200mm and 300mm wafers. Via resistance is an integrated measure of interconnect electrical performance. As the data show, 200mm and 300mm Via resistances are closely matched.

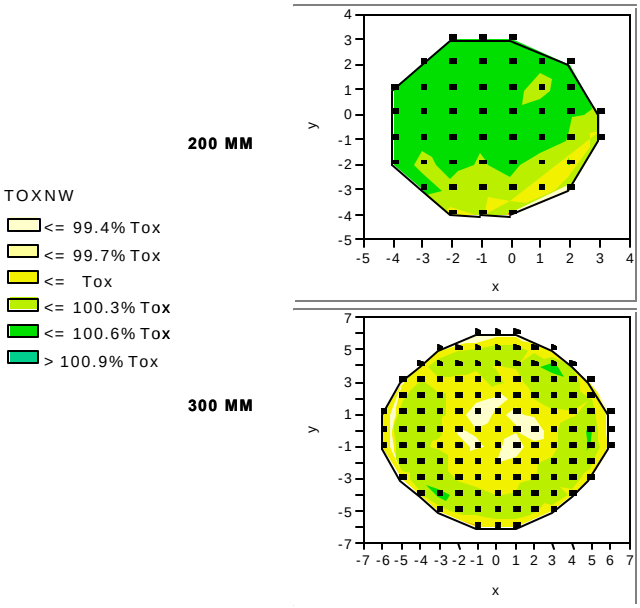


Figure 9: 200mm/300mm within-wafer gate oxide

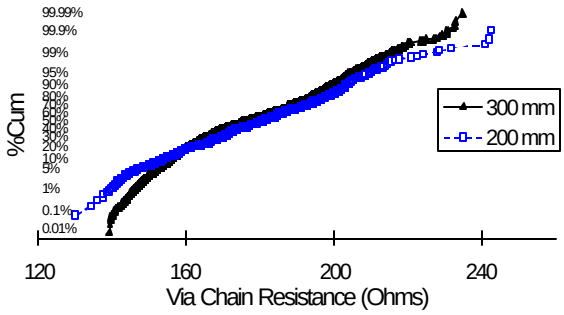


Figure 10: 200mm/300mm Via resistance distribution

Figures 11(a) and 11(b) show Transmission Electron Microscope (TEM) cross-sections of 200mm and 300mm gate electrodes. These are approximately identical, non-minimum gate-length transistors. Profiles and critical film thicknesses are well matched. Slight differences in the film conformality and interfaces are evident. These are unavoidable differences caused by configuration differences between the 200mm and 300mm tools.

Figure 12 shows a TEM cross-section of the complete 6-layer interconnect system. Profiles and thickness are virtually identical between 200mm and 300mm.

Pentium is a trademark of Intel Corporation or its subsidiaries in the United States and other countries.

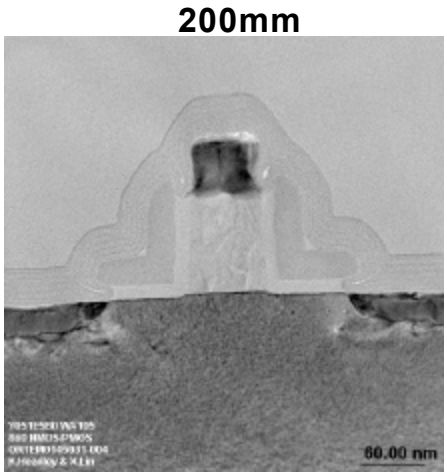


Figure 11(a): 200mm gate electrode TEMs

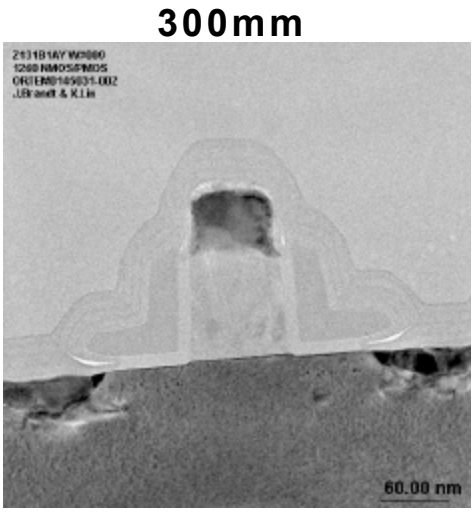


Figure 11(b) : 300mm gate electrode TEMs

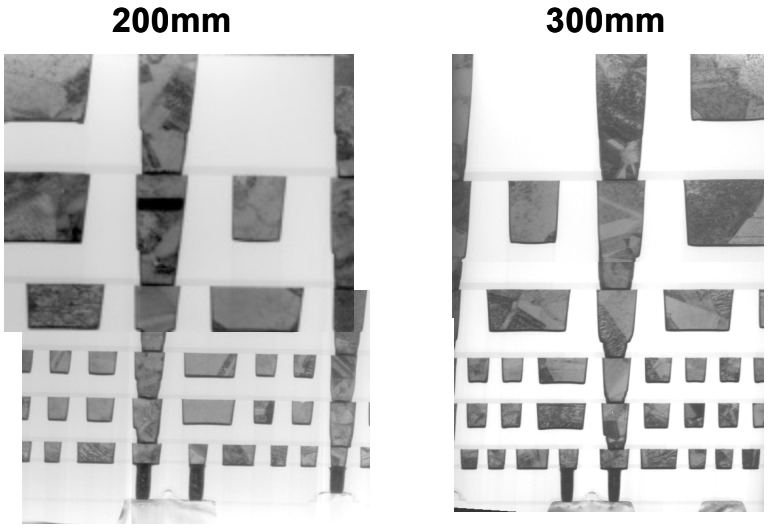


Figure 12: 200mm/300mm interconnect TEMs

Performance Matching

Figure 13 shows a basic transistor matching graph between 200mm and 300mm. Saturated drive current (Idsat) is plotted against off-state leakage (Ioff) for both 200mm and 300mm NMOS and PMOS transistors. The data show that the 200mm and 300mm devices are perfectly matched across a wide range of Ioff.

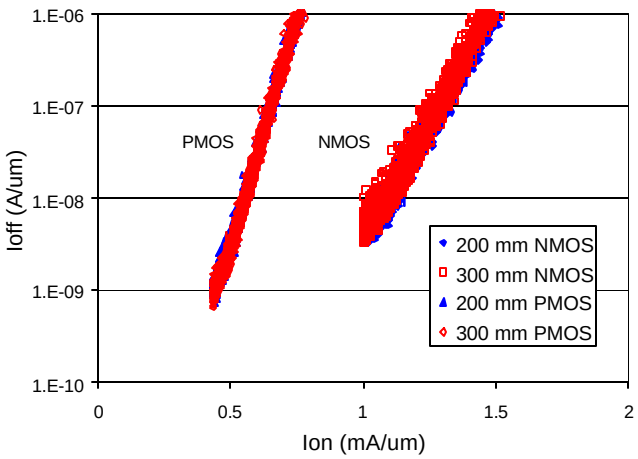


Figure 13: 200mm/300mm transistor Ion/Ioff

Figure 14 shows a circuit-level matching metric. The graph is a cumulative distribution of ring oscillator test circuit frequencies on 200mm and 300mm wafers. Again, the data indicate that the circuit operating frequencies are perfectly matched.

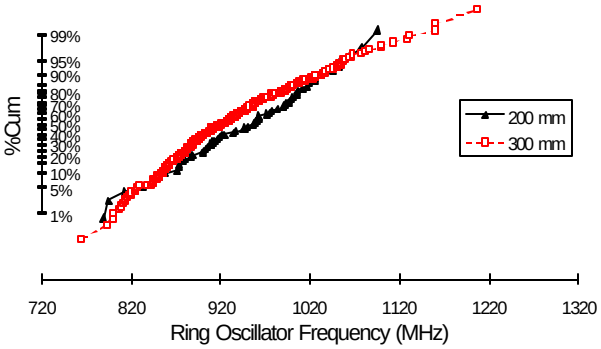


Figure 14: 200mm/300mm ring oscillator circuit frequency

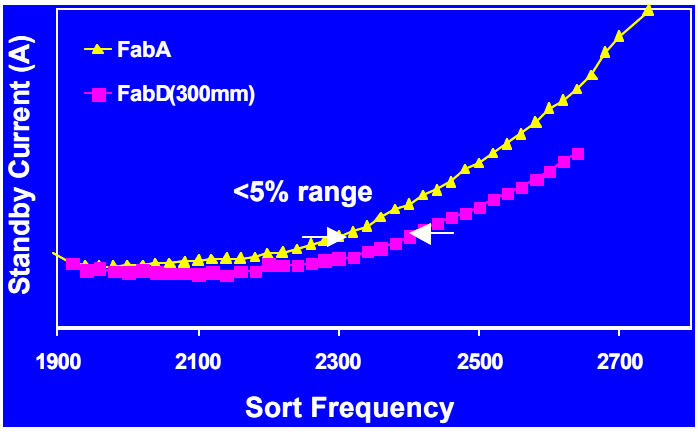


Figure 15: Pentium® 4 sort frequency vs. standby current

Finally, Figure 15 shows a normalized performance comparison for the Pentium® 4 product. The graph shows sort frequency graphed against product standby current. The 300mm product speed is within 5% of the reference 200mm population, matched to within normal variability.

Yield Matching

Figure 16 shows normalized die yield for 300mm and 200mm as a function of time. 300mm die yield at the start of development is lower than 200mm, which is shown starting after initial ramp. Rapid yield learning, facilitated by the ability to copy 200mm learning, enabled steadily improving die yields to the point where 200mm and 300mm die yields are matched at the point of the 300mm initial ramp.

Reliability Matching

Figure 17 shows a key transistor reliability metric, gate-oxide time-to-breakdown. The data are shown as a normalized distribution function of time-dependent dielectric breakdown (TDDB) in seconds. Both 200mm and 300mm are well matched in gate-oxide reliability.

Figure 18 shows a key interconnect reliability metric, electromigration fail rate. The data are shown as a normalized distribution function of time-to-fail. Again, both 200mm and 300mm are well matched.

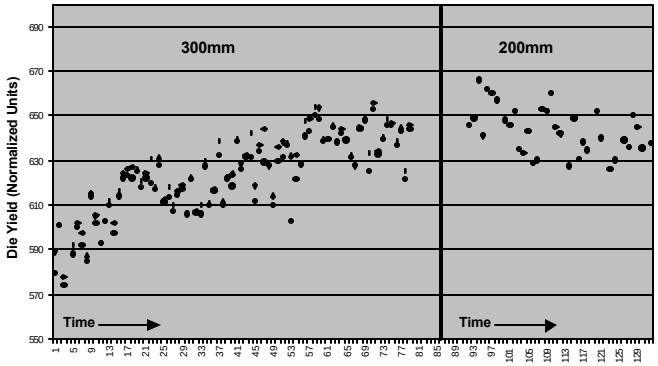


Figure 16 : 200mm/300mm normalize die yield

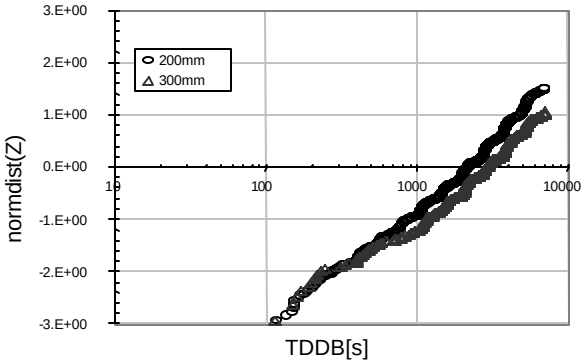


Figure 17: Gate oxide 200mm/300mm time-to-fail

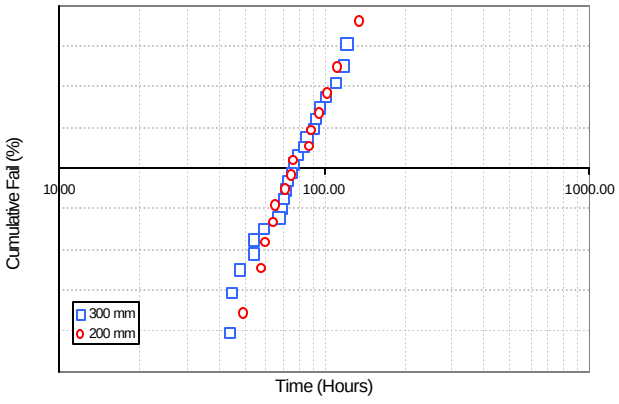


Figure 18: Electromigration cumulative fail rate

CONCLUSION

For over 35 years, Moore's Law has set a rapid pace for progress in the semiconductor industry. With the continuously increasing technical challenges for silicon technology development, increasingly rapid yield learning and volume manufacturing ramp rate have been instrumental in maintaining Intel's technology leadership.

In this paper, we discussed the implementation of Intel's industry-leading 0.13 m logic technology on the 300mm wafer size and associated process equipment. The 0.13 m process has been ramped to volume production in multiple factories and on both 200mm and 300mm production lines at record yields, quality, and ramp rate. Rapid development of Intel's first 300mm wafer-size technology, well matched to the 200mm state-of-the-art process, is a critical milestone for future competitiveness. The adaptation of proven Copy Exactly! methods is the key element that enabled successful conversion to the 300mm wafer size and sets the stage for Intel's continued leadership in the semiconductor industry.

ACKNOWLEDGMENTS

The authors thank the many outstanding engineers and technicians in the Logic Technology Development and 300mm OHT organizations who are responsible for developing, ramping, and transferring Intel's P1260 logic technology.

REFERENCES

[1] Tyagi, S., et. al., "A 130nm Generation Logic Technology Featuring 70nm Transistors, Dual Vt Transistors and 6 layers of Cu Interconnects," *IEDM Technical Digest*, December, 2000, pp. 567-570.

[2] Moore, G.E. "Cramming More Components Onto Integrated Circuits," *Electronics Magazine*, Vol. 8, April, 1965, pp. 114-117.

[3] Schaller, R.R. "Moore's Law: Past, Present, and Future," *IEEE Spectrum*, Vol. 34, Issue 6, pp. 52-59.

[4] Gasser, Jr., R.A., "Yield Learning and Volume Manufacturing of High-Performance Logic Technologies on 200mm and 300mm Wafers," *IEDM Technical Digest*, December, 2001, pp. 599-601.

AUTHORS' BIOGRAPHIES

Sanjay Natarajan is a process integration group leader in Intel's Logic Technology Development organization. He joined Intel in 1993 and has held numerous positions in both factory automation and process integration. He is presently responsible for transistor integration for Intel's 65nm CMOS process technology. Prior to this, he led transistor integration for Intel's first 300mm process technology, a 0.13 m CMOS process matched to Intel's 200mm technology. He received his B.S., M.S., and Ph.D. degrees from Carnegie Mellon University, all in Electrical Engineering. His e-mail is Sanjay.Natarajan@intel.com.

Melton Bost is a process integration group leader in Intel's Logic Technology Development organization. He joined Intel in 1987 and has worked primarily in back- end process integration since that time. He is presently responsible for the P1260 backend process technology. He received his B.S.E. degree from Duke University in 1978, his M.S.E. degree from Stanford University in 1979, both in Materials Science, and his Ph.D. degree from Colorado State University in 1987 in Electrical Engineering. He is the author of numerous technical papers and holds six patents. His e-mail is Melton.Bost@intel.com.

Derek Fisher is the P1260 Yield Group Leader in Intel's Logic Technology Development organization. Since 1994, he has been responsible for Defect Metrology Roadmaps and participated in Fab startup and process transfer activities from P854 to P1260. He joined Intel in Ireland in 1991 and worked on P652 to P852 transfer and yield improvement. Prior to that, he worked in manufacturing and process development roles for National Semiconductor, Philips Research, and Motorola. He graduated from Strathclyde University in 1983. His e-mail is Derek.G.Fisher@intel.com.

David Krick is the equipment startup coordinator for Intel's Logic Technology Development organization. He joined Intel in 1989 as a process engineer and has held numerous process engineering and management positions. He is currently responsible for coordinating the startup of Intel's newest 300mm technology development facility, D1D. Prior to this, he managed the successful startup of D1C, Intel's first 300mm factory. He holds three patents. Krick received his B.S. degree in Electrical Engineering and his M.S. degree in Engineering Science, both from Pennsylvania State University. His e-mail is David.T.Krick@intel.com.

Chris Kenyon is a lithography group leader in Intel's Logic Technology Development organization. He joined Intel in 1996 and has worked primarily on Intel's gate patterning process since that time. He is currently responsible for developing the gate patterning process for the 90nm CMOS node. He received his B.A. degree from Princeton University in 1990 and his Ph.D. degree from Caltech in 1996 in Physical Chemistry. His e-mail is Chris.Kenyon@intel.com.

Chris Kardas is a process engineering group leader in Intel's Logic Technology Development organization. Since joining Intel in 1984, he has worked on a variety of Etch modules where he holds a patent. Most recently he has been working in the area of lithography and is presently responsible for P1260 and P1212 front-end layers. He received his B.S.E.E. degree from the University of Illinois in 1984. His e-mail is Chris.Kardas@Intel.com.

Chris Parker is a senior process engineer in Intel's Logic Technology Development organization. He joined Intel in 1998 and has worked on front-end oxidation and gate development. He is presently responsible for the development of alternative gate dielectric processes for 300mm P126x technologies. He received his B.S.E.E. degree from Auburn University and M.S. and Ph.D. degrees in Electrical Engineering from North Carolina State University. His e-mail is Chris.PTD.Parker@intel.com.

Robert A. Gasser, Jr. is Vice President, Technology & Manufacturing Group, and Director, Components Research. He is responsible for research and development of process technologies used to build future Intel logic devices that will be in production five to ten years from now. He joined Intel in 1982 as a Technology Evaluation Engineer. Most recently, Gasser was responsible for the development of Intel's 0.18-micron logic process technology (P858) and 300mm, 130nm logic technology (P1260) and 90nm logic technology (P1262). Gasser received his B.A. degree in Physics from Reed College in 1980. He received his M.S. degree in Materials Science from Stanford University in 1982. Gasser has written numerous technology papers and holds five patents. His e-mail is Bob.Gasser@intel.com.

Copyright © Intel Corporation 2002. This publication was downloaded from http://developer.intel.com/.

Legal notices at http://www.intel.com/sites/corporate/tradmarx.htm.

ETOX™ Flash Memory Technology: Scaling and Integration Challenges

Al Fazio, California Technology and Manufacturing, Intel Corp.
Stephen Keeney, California Technology and Manufacturing, Intel Corp.
Stefan Lai, California Technology and Manufacturing, Intel Corp.

Index words: Flash memory, ETOX™, Intel StrataFlash® memory, Moore's Law

ABSTRACT

The 0.13 m flash memory technology that started high-volume manufacturing in the first quarter of 2002 is the eighth generation of flash technology since its first conception and development in 1983. The scaling has been accomplished by improved lithography capability as well as many process architecture innovations. In this paper, the key scaling challenges as well as the key innovations are presented. It is projected that the current planar cell structure can be scaled to the 65nm node. More revolutionary innovations, such as 3D structures, may be required for the 45nm node and beyond. To lower cost further, Intel StrataFlash memory technology has been developed, which stores two bits of information in a single physical memory cell. The scaling innovations also allow for the integration of flash memories with high-performance logic for "wireless Internet on a chip" technology. These integration challenges are also discussed.

INTRODUCTION

The in-system update and non-volatile capabilities of flash memories have enabled it to become the memory of choice for many emerging markets over time, originally as point of sales system configurations, then as PC BIOS components, and today for cell phones and handheld computing devices [1]. Similar to other memory technologies, ETOX™ flash memory scaling follows Moore's law. Figure 1 shows SEM cross-sections of the memory cells for eight generations of flash memory technologies. The memory cell size for the first generation based on 1.5 m lithography was 36 m^2, whereas the cell size for the latest 0.13 m lithography is 0.154 m^2. This represents an over 230 times cell size reduction over the eight generations. In the same period, the memory density for peak volume has increased one thousand fold from 64Kb to 64Mb.

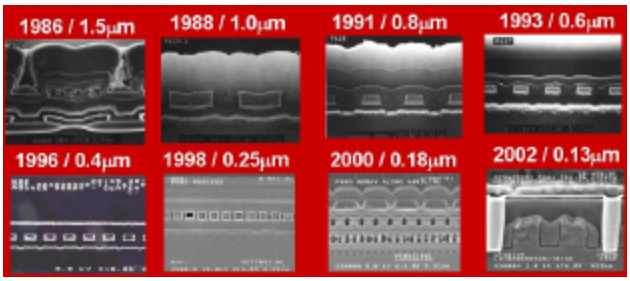


Figure 1: Eight generations of flash technology

Although scaling the flash cell is important to achieve die size reduction or larger memories, the periphery transistors must also be scaled. Scaling the periphery transistors can be achieved by reducing the maximum voltages that need to be supported along with junction engineering and more advanced lithography and etch capabilities. The process architecture innovations and scaling of periphery transistors enables the integration of flash memories with high-performance logic for "wireless Internet on a chip" technology. In this paper we review the key process architecture innovations for scaling, the Intel StrataFlash memory technology and the key innovations required for "wireless Internet on a chip" technology. Table 1 outlines the key innovations for each generation of flash memory.

StrataFlash and ETOX are trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Table 1: Innovations by technology generation

Technology Node	Key Innovation
1.5 m	Established Flash
1.0 m	Isolation rounding reduction for improved cell gate alignment Cycling reliability established
0.8 m	Recessed LOCOS
0.6 m	Self Aligned Source Scaled Array Field Oxide
0.4 m	Negative Gate Erase Intel StrataFlash memory
0.25 m	Trench Isolation Salicide
0.18 m [2]	Self Aligned floating gate Unlanded Contacts Multiple Periphery Gate Oxides
0.13 m [3]	Channel Erase Dual Trench Dual gate Spacer Wireless Internet on a Chip

FLASH CELL SCALING

Cell size scaling is achieved by scaling critical area components. Each of the key scaling components is described. Figure 2 illustrates cell layout and scaling constraints. A key enabler to scaling is improved line width and space definition through new lithography at each generation. Architecture innovations, such as a number of self-aligned techniques, provide the bulk of the remaining area reduction.

CELL WIDTH (WORDLINE DIRECTION)

The cell width is determined by the simultaneous constraints of isolation pitch (isolation and cell active diffusion); floating gate pitch (endcap, space, and alignment); and contacted metal pitch (contact size, contact and metal space, and alignment). Each of these needs to be scaled in order to scale the cell width.

Isolation Pitch

Two key approaches have been adopted over the last several generations that have enabled continuous pitch scaling. The first is the adoption of a dual isolation scheme where the flash array isolation is decoupled from the periphery isolation so each can be optimized independently. This was first introduced in a local oxidation of a silicon isolation scheme, LOCOS, in the 0.6 m generation. The second key enabler was the introduction of trench isolation at the 0.25 m node, which helped to reduce the active width loss in the device. For the 0.13 m generation, a dual isolation scheme was adopted, now called dual trench, where the array trench was made shallower than the periphery trench for independent optimization. As before with the dual LOCOS scheme, the flash cell can be scaled more aggressively while still meeting the periphery isolation requirements. At each technology node, improved lithography capability is utilized. Additionally, improved gap fill capability of High-Density Plasma (HDP) oxides has been utilized since the 0.18 m technology node.

Floating Gate Pitch

The correct alignment of the floating gate to the active area is a very important cell size determinant, and it becomes more of a constraint as the isolation pitch is scaled and the floating gate isolation is constrained by the lithography minimum space capability. The 0.18 m technology node introduced a new self-aligned scheme (Figure 3, left half) where the floating gate is self-aligned to the isolation using a chemical mechanical polish process. This has been carried forward to the 0.13 m node as well. This self-aligned scheme removes the registration component of the scaling and also allows a sub-lithographic poly space.

Contacted Metal Pitch

Each generation takes advantage of the advances in lithography to scale the contact size and metal pitch. However, the contact alignment to the active area became the constraint at the 0.18 m node, and an UnLanded Contact (ULC) scheme was introduced (Figure 3, right half). In this case, a nitride etch stop layer is deposited below the inter-layer-dielectric oxide to prevent the contact etch punching through the isolation and causing a short to the substrate. This allows the contact to land partially in the isolation and reduces the registration constraint. This ULC scheme is continued in the 0.13 m technology.

CELL HEIGHT (BITLINE DIRECTION)

The cell height is determined by constraints of contact size and contact-to-gate alignment, gate length and drain and source space (source rail width).

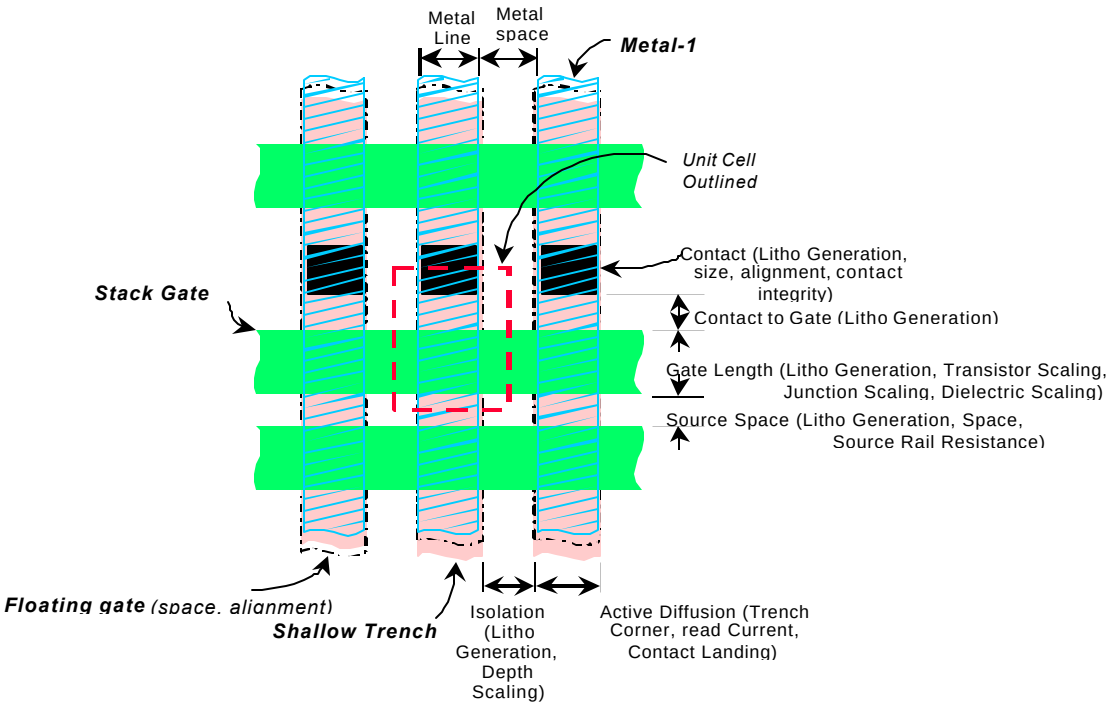


Figure 2: Cell layout and scaling constraints

Contact Size

The key determinants to contact scaling have been the advances in lithography tools, resists, and masks. These have enabled the printing of smaller contacts at every generation. This has been coupled with advances in contact etch chemistry along with the adoption of salicided junctions starting at the 0.25 m generation, eliminating the need for plug implants, required by non-salicided contact processes. The contact plug uses PVD Ti/CVD TiN adhesion layers and blanket tungsten deposition followed by chemical-mechanical polish. The unlanded contact process introduced at 0.18 m (Figure 3, right half) improved registration by allowing a direct contact-to-gate alignment without worrying about alignment to the isolation.

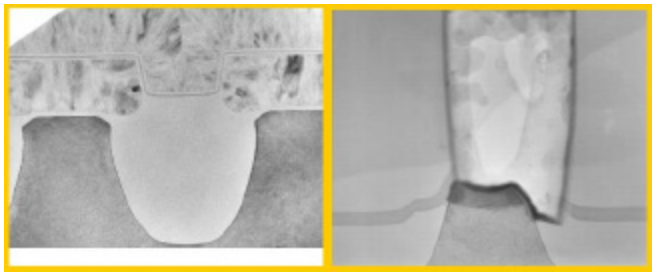


Figure 3: Self-aligned poly and unlanded contact

Source Space Scaling

The primary challenge to scale the source space is to meet the source resistance requirements for each generation. Similar to the contact, the most advanced lithography is used to define the poly space at each generation. A self-aligned source architecture was introduced in the 0.6 m node to eliminate the registration component of the flash cell gate to the diffusion edge, and this continues to be used today. To prevent the source resistance from increasing beyond the maximum requirement, the trench profile and source implants are carefully engineered to manage the trench sidewall resistance without the need for angled implants. The adoption of a dual trench scheme in the 0.13 m generation allowed a much shallower trench to be chosen for the flash array, which made it easier to dope the sidewall, especially at the tighter pitch.

Gate Length Scaling

Gate length has been scaled at each generation using similar techniques to classical transistor scaling, which include junction and channel doping optimization along with gate oxide scaling. In the case of flash, both the tunnel oxide thickness and the interpoly Oxide-Nitride-Oxide (ONO) thickness are scaled to improve the gate coupling to the channel to allow further channel length scaling. In the 0.13 m generation, the ONO effective electrical thickness is 15nm, and the tunnel oxide thickness is 9nm. Changes to the erase scheme have also aided in channel length scaling by allowing the source junction to be scaled, thereby reducing the source junction underlap. At the 0.4 m generation a negative gate erase scheme was adopted, which reduced the cell source voltage from 12V

in the source erase scheme used in earlier generations to ~5V with negative gate erase. At the 0.13 m generation a channel erase scheme was adopted so that the junction could be scaled further as it now no longer needs to support a voltage above the well voltage.

Drain Space Scaling

Generally the drain space is not limited by lithography, as it is larger than the source space, due to the presence of a contact. The key concerns with drain space scaling are adequate contact-to-gate space, which is reduced with improved registration, Inter-Layer-Dielectric gap fill (a HDP oxide is used at the 0.18 m generation and beyond), and the spacer architecture. In the transition from 0.18 m to 0.13 m, a dual spacer scheme was adopted that allowed the flash array spacer to be independent from the periphery high-voltage transistors. This enabled a narrower spacer in the flash drain region so that gap fill was not an issue.

SCALING LIMIT PROJECTION

One can extrapolate the scaling trend based on what has been accomplished so far and the result is shown in Figure 4. This extrapolation is based on the fact that the basic planar cell structure is the same for all the generations, and scaling is achieved by reducing specific cell dimensions. The active electrical cell area is Zeff x Leff, which represents the minimum area required for cell functionality. The trend was relatively flat from 1.0 m to 0.40 m nodes, but was scaled aggressively since the 0.25 m generation. Zphy and Lphy represent the active width and gate length dimensions defined lithographically. The difference between Zphy and Zeff is the beak of the isolation process while the difference between Lphy and Leff is the lateral diffusion of the source and drain underneath the gate. Zphy x Lphy is scaling down at a faster rate compared to Zeff x Leff because of the aggressive reduction of beak and source/drain underlap. However, the beak and source/drain underlap cannot go to zero. Thus, the convergence point of the trend represents a projection of a scaling rate limiter of the current planar cell structure. The trend shows convergence at 45nm, which means that this component of scaling is no longer available. A practical limit of scaling of this component is the 65nm node. This also agrees well with analyses based on other considerations. To continue scaling at the same rate, i.e., meeting Moore's Law, more revolutionary ideas will be needed to either scale the Leff and Zeff more aggressively, which is historically difficult due to hot electron programming limitations, or to go to other cell structures that are not planar (3D cell structures).

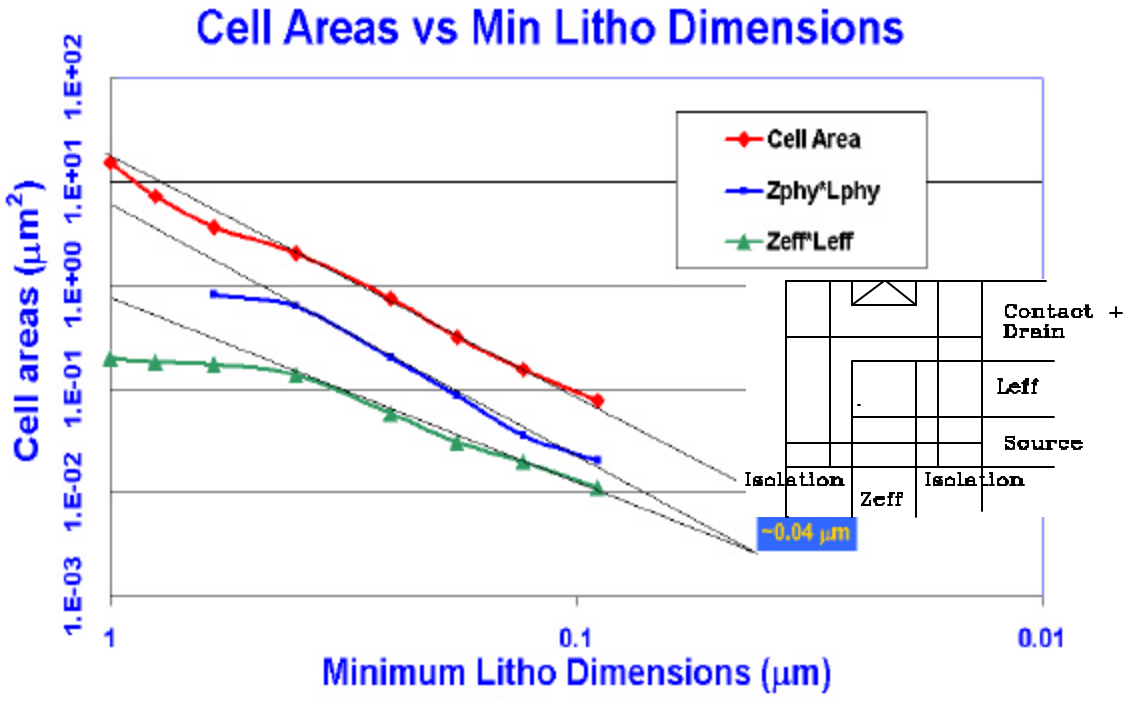


Figure 4: Cell scaling projection

INTEL STRATAFLASH® MEMORY

The Intel StrataFlash memory technology represents a cost breakthrough for flash memory devices by enabling the storage of two bits of data in a single flash memory transistor. Cost-per-bit reduction of flash memory devices has been traditionally achieved by aggressive scaling of the memory cell transistor using silicon process-scaling techniques as discussed in the previous sections of this paper. In an attempt to accelerate the rate of cost reduction beyond that achieved by process scaling, a research program was started in 1992 to develop methods for the reliable storage of multiple bits of data in a single flash memory cell. The result of this research was the commercial introduction of the first Intel StrataFlash memory in 1997, utilizing the 0.4 m technology node. The two-bit-per-cell Intel StrataFlash memory technology provides a cost structure equivalent to the next generation of process technology while using the current generation of process technology equipment. Today, the Intel StrataFlash memory technology has become the mainstream flash solution.

The Multi-Bit Storage Breakthrough: Intel StrataFlash® Memory Technology

As discussed earlier, the flash memory device is a single transistor that includes an isolated floating gate. The floating gate is capable of storing electrons. The behavior of the transistor is altered depending on the amount of charge stored on the floating gate. Charge is placed on the floating gate through a technique called programming. The programming operation generates hot electrons in the channel region of the memory cell transistor. A fraction of these hot electrons gain enough energy to surmount the 3.2eV barrier of the Si-SiO_2 interface and become trapped on the floating gate. For single-bit-per-cell devices, the transistor either has little charge (<5,000 electrons) on the floating gate and thus stores a "1," or it has a lot of charge (>30,000 electrons) on the floating gate and thus stores a "0." When the memory cell is read, the presence or absence of charge is determined by sensing the change in the behavior of the memory transistor due to the stored charge. The stored charge is manifested as a change in the threshold voltage of the memory cell transistor. Figure 5 illustrates the threshold voltage distributions for a half-million cell (1/2Mc) array block. After erasure or programming, the threshold voltage of every memory cell transistor in the 1/2Mc block is measured, and a histogram of the results is presented. Erased cells (data 1) have threshold voltages less than 3.1v, while programmed cells (data 0) have threshold voltages greater than 5v.

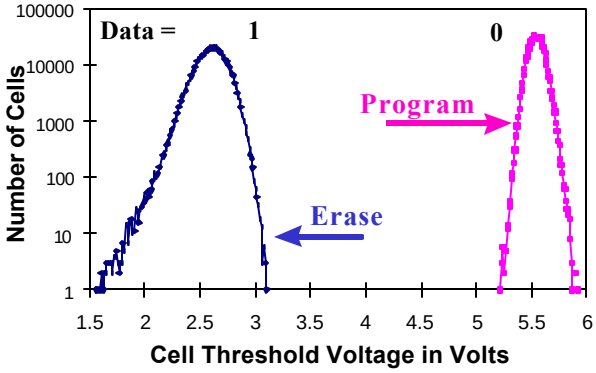


Figure 5: Single-bit/cell array threshold voltage histogram

The charge storage ability of the flash memory cell is a key to the storage of multiple bits in a single cell. The flash cell is an analog storage device, not a digital storage device. It stores charge (quantized at a single electron), not bits. By using a controlled programming technique, it is possible to place a precise amount of charge on the floating gate. If charge can be accurately placed to one of four charge states (or ranges), then the cell can be said to store two bits. Each of the four charge states is associated with a two-bit data pattern. Figure 6 illustrates the threshold voltage distributions for a 1/2Mc block for two bits per cell storage. After erasure or precise programming to one of three program states, the threshold of each of the 1/2Mc is measured and plotted as a histogram. Notice the precise control of the center two states, each of which is approximately 0.3v (or 3,000 electrons) in width.

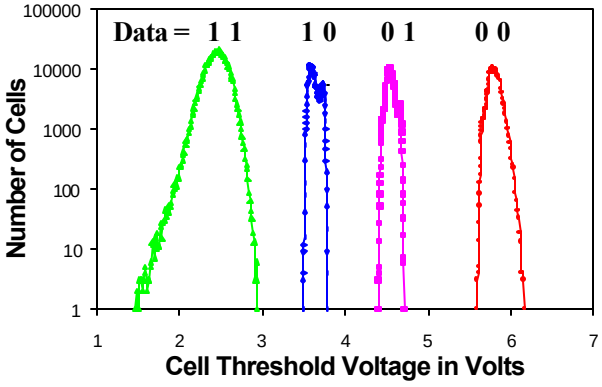


Figure 6: Two-bit/cell array threshold voltage histogram

Higher bit-per-cell densities are possible by even more precise charge placement control. Three bits per cell require eight distinct charge states and four bits per cell

StrataFlash is a registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

require sixteen distinct charge states. In general, the number of states required is equal to 2^N where N is the desired number of bits.

The ability to precisely place charge on the floating gate and at some later time sense the amount of charge that was stored has required substantial innovations, and extensive characterization and understanding of cell device physics, memory design, and memory test. These innovations are discussed in detail in two earlier *Intel Technology Journal* papers [4,5].

LOW-VOLTAGE, HIGH-PERFORMANCE OPTIMIZATION FOR DISCRETE FLASH MEMORIES

Increasing read performance demands at low operating voltage taxes the ability of high-voltage transistors, which are required by flash program and erase. Cobalt-salicided complementary polysilicon gates are used to form low-threshold NMOS and PMOS surface-channel transistors and low source/drain and gate resistance. Additionally, special low-threshold devices, for low-voltage performance and analog circuit design requirements, are provided by separate well and Vt-adjust implants. Continual application performance demands and further reductions in operating voltages require the inclusion at the 0.18 m technology generation of thin gate-oxide logic compatible NMOS and PMOS transistors. This is achieved with three additional masking layers (one for thin gate oxide and two for low-voltage wells). Source/drain and tip regions are shared between the low-voltage and high-voltage transistors to best balance performance with added processing steps. The thin gate-oxide architecture is bounded by optimization for low voltage (<1.8V), while maintaining compatibility for legacy voltage (3.3V), including balancing of device Vt with off-current leakage for minimization of standby currents. An 8nm gate oxide was chosen to balance these needs, with trench processing meeting charge-to-breakdown requirements, supporting three separate oxides: tunnel oxide, high-performance oxide, and high-voltage oxide. A triple well is provided for design flexibility of negative voltage switching and low-voltage optimization. Lastly, performance capability is provided by three layers of aluminum metalization, allowing additional wordline and bitline strapping of the flash array, for reduced resistance-capacitance, RC delay, and more efficient signal routing in the periphery.

In addition to low voltage and high performance, the trench isolation, thin-gate-oxide, salicided complementary poly gate transistors and the three layers of interconnect inclusion provide all the key architectural elements required for embedded logic capability. Higher degrees of thin-gate device performance can be achieved by further separating the process steps and reducing the oxide thickness for lower voltage operation, as discussed below.

Lastly, the cost sensitivity of the market for memory dictates requirements for low-cost process technologies. The described cell scaling and Intel StrataFlash memory capability satisfy low cost. Additionally, process synergy of this memory process technology, with the basic process modules and equipment set with other high-volume logic technologies, lower cost through economies of scale by providing factory flexibility and shared process step and yield learning.

To reduce cost, the periphery transistors must also be scaled since they constitute a significant portion of the die area. The introduction of channel erase reduces the maximum voltage the periphery needs to support, and the introduction of more advanced lithography and etch gives better gate-patterning capability. These allow the channel length and gate oxides to be scaled, which is done in conjunction with traditional junction scaling, and which leads to a significant reduction in the gate length, while at the same time maintaining good transistor characteristics. For the embedded logic process, below, this leads to a gate length of 100nm. The reduction in the maximum voltage the periphery needs to support along with the dual trench scheme allows the isolation width to be scaled as well, since a deep trench can be maintained for logic devices independent of the shallow trench used in the flash array. These changes, combined with the advanced 0.13 m lithography tools, cobalt salicide, and complimentary gates consistent with Intel's 0.13 m logic process, deliver the required transistor performance and area savings.

WIRELESS INTERNET ON A CHIP

Traditionally, flash and logic process technologies are optimized separately on separate process equipment sets and separate fabrication facilities. During the development of the 0.25 m flash process technology, Intel made the strategic decision to develop its flash processes synergistically with its logic processes. This initial decision was made with the goal of processing the two technologies, flash and logic, in the same fabrication facility, for improved manufacturing flexibility and shared learning and for maximum volume production efficiency. This decision also brought key process modules into the flash processes, which historically were not found on flash, such as trench isolation and salicided

StrataFlash is a registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

complimentary gates. Both of these process modules are examples of key enablers that not only achieve the manufacturing synergy goal, but also provide for dramatic advancement of scaling the flash memory cell and enhancing performance. (This was outlined in the previous sections.) Additionally, the incorporation of these features into flash memory technology has paved the way for the integration of a high-performance logic function with a dense flash memory on the same chip. This capability has led to the "wireless Internet on a chip" technology, where all the key elements of a typical cell phone and a typical handheld computer, the advanced digital logic functions, all the SRAM and flash memory functions, and the analog functions for interfacing to a radio are all integrated onto a single chip. This is cost-effectively achieved without compromising the performance of the state-of-the-art digital logic or the density of the state-of-the-art flash memory.

The value of this integration is several fold. First, the total number of devices can be reduced, thereby reducing the form factor of a wireless device, allowing for smaller lighter devices. What were previously several chips is now reduced to one. The reduction in the number of chips in a system also improves overall system reliability. Secondly, the integration of flash memory serves to enhance the performance of the digital logic computing functions. Memory latency is greatly reduced, and bandwidth is greatly enhanced by having logic and memory functions integrated onto the same chip. Lastly, this enhanced performance is achieved at lower power, as interconnect bus capacitance is significantly reduced with an integrated on-chip bus, versus a discrete external bus.

Five key innovations are required to achieve the "wireless Internet on a chip" technology. They are the key process modules for advanced logic functionality with an advanced flash process. These five innovations are trench isolation, a multiple gate-oxide process, a low thermal budget, salicided complementary gates, and a multi-level metal system.

Trench isolation is required for tight pitch logic design rules, for high transistor count design, and for small SRAM memory cell layout. Trench isolation is not typically found on flash processes, due to the challenges outlined earlier. These challenges were overcome with the integration of trench isolation in the 0.25 m flash process and have served as the basis for cell size reduction in the flash cell.

Multiple gate oxides are required to achieve the separate function required for the high-voltage operation of the flash cell and the ultra-low voltage required for the logic operation. The 0.18 m flash process incorporated multiple periphery gate oxides, as outlined earlier. This same process architecture is extended to achieve the ultra-thin (<3nm) gate oxide required for advanced logic functions.

A low thermal budget processing is required for high-performance transistors. Traditionally, memory plus logic integrated with memories such as DRAMs have had difficulties with achieving a low thermal budget, as the DRAM cell processing (requiring high temperatures) is often done subsequent to the formation of the logic transistors, thereby significantly limiting the performance of the logic functions. This is not the case with flash memory integration, as the flash memory processing occurs earlier than the formation of the logic transistors. As such, the high-thermal process steps of the flash memory are not seen by the logic transistors, thereby maintaining the high-performance capability of the logic functions.

Salicided complementary gates are required to achieve low-threshold voltages and short channel lengths that are required for high-performance logic functions. Salicided gates are often difficult to integrate with memories, as tight spaces found in memories pose challenges to salicide processing. These barriers were overcome in the 0.25 m node flash technology, with the integration of salicide, outlined earlier.

Lastly, multiple metal layers are required for high transistor-count logic designs. The metal processing is accomplished with backend planarization, fully compatible with the logic and flash processing.

With these innovations, the ability to fully integrate state-of-the-art logic performance and state-of-the-art flash memory density, cost effectively, without compromising either, has been fully realized. The analog features are relatively simple process components, most of which are found in standard flash processing. Key attributes for analog processing are a triple well for noise isolation that is standard in flash memories, 3V optimized transistors that are also standard in flash memories, and precision resistor and capacitor passives, that can be bolted on, relatively simply, to a CMOS process.

CONCLUSION

By following Moore's law, ETOX™ flash memory has gone from 1.5 m node in development in the mid 1980s to 0.13 m in high-volume production today. The scaling has been accomplished by improved lithography capability as

ETOX and StrataFlash are trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

well as many innovations. In this paper, we reviewed key scaling challenges as well as the key innovations. Based on projection, the current planar cell structure can be scaled to the 65nm node. More revolutionary innovation such as 3D structures may be required for the 45nm node and beyond. To lower cost further, we have developed the Intel StrataFlash® memory technology, which stores two bits of information in a single physical memory cell. The scaling capability also allows for the integration of flash memories with high-performance logic for "wireless Internet on a chip" technology.

ACKNOWLEDGMENTS

The authors acknowledge the members of Intel's California Technology and Manufacturing organization and the Flash Products Group for their efforts over the past eight generations of technologies.

REFERENCES

[1] S. Lai, "Flash Memories: Where We Were and Where We Are Going," *IEEE IEDM Tech. Digest*, 1998, pp. 971-3.

[2] A. Fazio, "A High Density High Performance 180nm Generation High Density Etox™ Flash Memory Technology," *IEEE IEDM Tech. Digest*, 1999, pp. 267-270.

[3] S. Keeney, "A 130nm Generation High-Density Etox™ Flash Memory Technology," *IEEE IEDM Tech. Digest*, 2001, pp. 41-44.

[4] G. Atwood, et. al., "Intel StrataFlash Memory Technology Overview" *Intel Technology Journal*, Q4, 1997 at http://developer.intel.com/technology/itj/q41997/articles/art_1.htm

[5] A. Fazio, et. al., "Intel StrataFlash Memory Development and Implementation" *Intel Technology Journal*, Q4, 1997 at http://developer.intel.com/technology/itj/q41997/articles/art_2.htm

AUTHORS' BIOGRAPHIES

Al Fazio is a Principal Engineer responsible for Communication Technology Development. He joined Intel in 1982 after receiving his B.S. degree in Physics from the State University of New York at Stony Brook. He has worked on numerous memory technologies and was responsible for the Intel StrataFlash memory and Flash+Logic+Analog "wireless Internet on a chip" technology developments. Al holds over 20 patents and has written several technical papers, two of which have won outstanding paper awards at IEEE-sponsored conferences. He has served as general chairman of the IEEE NVSMW. His e-mail is al.fazio@intel.com

Stephen Keeney is the Process Integration Manager for the 0.13 m and 0.09 m flash technology development programs. Stephen obtained his B.E. degree from University College Dublin, Ireland in 1988 and his Ph.D. degree in Microelectronics from the NMRC, Cork, Ireland in 1992. He joined Intel in 1993 and has worked extensively across many aspects of flash memory development, including device physics innovations; yield analysis, memory test architecture, process integration and Intel StrataFlash memory. Stephen holds six patents and has written 20 technical papers. His e-mail is stephen.n.keeney@intel.com

Stefan K. Lai is Vice President, Technology and Manufacturing Group, and Director, California Technology and Manufacturing. Stefan is responsible for the development of silicon process technologies for devices used in communications products, including flash, flash+logic, analog, and novel memory technologies. He was recognized as an IEEE Fellow in 1998 for his research on the properties of silicon MOS interfaces and the development of flash EPROM memory. Stefan received a B.S. degree in Applied Physics from the California Institute of Technology in 1973, and a Ph.D. degree in Applied Quantum Physics from Yale University in 1979. He joined Intel in 1982. His e-mail is stefan.lai@intel.com

Copyright © Intel Corporation 2002. This publication was downloaded from http://developer.intel.com/.

Legal notices at http://www.intel.com/sites/corporate/tradmarx.htm.

Integration of Mixed-Signal Elements into a High-Performance Digital CMOS Process

Kelin J. Kuhn, Technology and Manufacturing Group, Intel Corporation
Shahriar Ahmed, Technology and Manufacturing Group, Intel Corporation
Peter Vandervoorn, Technology and Manufacturing Group, Intel Corporation
Anand Murthy, Technology and Manufacturing Group, Intel Corporation
Borna Obradovic, Technology and Manufacturing Group, Intel Corporation
Kartik Raol, Technology and Manufacturing Group, Intel Corporation
Wai-kai Shih, Technology and Manufacturing Group, Intel Corporation
Iwen Chao, Intel Communications Group, Intel Corporation
Ian Post, Technology and Manufacturing Group, Intel Corporation
Steve Chambers, Technology and Manufacturing Group, Intel Corporation

Index words: RF-CMOS, HBT, inductor

ABSTRACT

The rapid increase in Internet communications' products such as high-speed switches, SerDes (serial-deserializer) elements and XAUI (X=10G, attachment unit interface) ports has energized the need for process technologies that support both digital and analog (mixed-signal) elements at radio frequencies (RF). In order for these products to be competitive, process technologies that support analog/mixed-signal and RF must heavily leverage the manufacturing benefits of conventional high-speed digital CMOS processes.

This paper reviews the challenges encountered when extending a high-speed conventional digital CMOS process to include analog/mixed-signal elements operating at RF frequencies.

INTRODUCTION

Analog/mixed-signal/RF product designs incorporating both CMOS and Bipolar Junction Transistor (BJT) active elements have emerged as a potential growth technology for the Internet communications marketplace.

Unlike older BiCMOS designs, these modern designs are adopting approaches where CMOS digital cells from well-characterized libraries are being mixed with specialized analog BiCMOS (or bipolar) modules. This enables rapid design modification, but places more demands on the process to deliver simultaneously optimized digital and analog elements [1].

Another key aspect of the analog/mixed-signal/RF process is the presence of passive elements. These passive elements include process-enabled elements (such as resistors and vertical capacitors) as well as design-enabled elements (such as inductors, varactors, and lateral capacitors).

SCALING CMOS (AND BICMOS) ACTIVE ELEMENTS

Digital CMOS has been successfully scaled for many generations and the performance optimization principles are well understood. Industry roadmaps such as the Semiconductor Industry Association (SIA) roadmap [2] and traditional industry scaling literature [3,4,5,6] reflect a long tradition of CMOS optimization.

In contrast, the optimization roadmap for analog/mixed-signal/RF is more difficult, due to conflicting digital and analog needs. The first problem is lack of commonality between digital and analog optimization strategies. The second problem is that desired analog process optimization strategies may run counter to traditional CMOS scaling, thus putting the extended analog/mixed-

signal/RF processes at odds with their mainstream CMOS progenitors.

The first problem (lack of commonality between digital and analog optimization priorities) can be more clearly illustrated by Tables 1 and 2. The tables represent an ordered list of CMOS DC parametrics and their importance in either analog or digital optimization. (The first item in each list is considered the most important.)

As just one example, note that I_{off} (of great priority to the digital designer) is not highly prioritized by the analog designer.

Table 1: Digital optimization strategy for DC parametrics

MOS - digital			
Parameter	Type	Units	Desired
Idsat	DC	mA/um	increase value
Ioff	DC	nA/um	decrease value
Vdd	DC	volts	decrease value
Vt	DC	mV	100mv < Vt < 300mV
Igate	DC	nA/um^2	decrease value

Table 2: Analog optimization strategy for DC parametrics

MOS - analog/RF			
Parameter	Type	Units	Desired
Vt	DC	mV	100mv < Vt < 300mV
gm	DC	uA/V	increase value
gds	DC	uA/V	decrease value
matching	DC	%	decrease differences
Igate	DC	nA/um^2	decrease value
Vcc	DC	volts	increase value
Ioff	DC	nA/um	decrease value

The second problem (desired analog process optimization strategies may run counter to traditional CMOS scaling) is becoming increasingly apparent in the literature. An illustrative example is I_{on}/I_{off} versus g_m/g_{ds}. Traditional CMOS scaling methodologies incorporate halo (pocket) implants to control short channel effects. However, halos have a detrimental effect on g_m/g_{ds} due to drain bias-induced modulation of the barrier created by the halo on the drain side of the device. This is a well-known problem and strategies ranging from lateral workfunction grading [7] to asymmetric halos [8] have been proposed. However, each of these strategies further removes the devices from the base technology, adding cost and complexity to the process.

Models and Measurements

A very critical part of process development is the ability to rapidly and precisely measure devices during the process development cycle and use these data to construct accurate and predictive device models. This creates both measurement and modeling challenges for the analog/mixed-signal/RF processes.

On the measurements side, although digital parts are well known to run at multi-GHz frequencies, CMOS digital optimization strategies do not require routine evaluation of RF metrics as part of process development. An analog/mixed-signal/RF process must enable such routine evaluation in order to produce accurate models at the 10+GHz of competitive products.

Since the digital community rarely evaluates RF metrics, the traditional metrics of the RF community become the metrics by default. These can be summarized as cut-off frequency (f_T), maximum oscillation frequency (f_{max}), minimum noise figure (NF_{min}) and noise figure at 50-ohm (NF_{50}), linearity (V_{IP3}), and 1/f noise level (usually shown as spectral noise density S_{vgate}) [1, 9,10].

Significant enhancements are required in CMOS measurement as well as in test chip design to support manual and automated RF measurements. More specifically, RF measurements require understanding and implementation of sophisticated de-embedding strategies [11]. RF devices are exceptionally sensitive to subtle differences in geometry [12], and test chip designs must incorporate significantly more device-specific calibration structures.

On the modeling side, circuit-level device models must clearly deliver accurate predictions at increasingly high frequencies. In addition, accurate simulators that enable noise figure calculation, noise parameter characterization (NF_{min}, G_{opt}, B_{opt}, and R_n) and time-domain noise simulation are indispensable to designers making trade-offs between power transfer and noise reduction [13].

Conventional compact models used in simulating digital circuits (e.g., BSIM3v3) lose accuracy at RF frequencies because the parasitic effects on high-frequency signals as they travel down the extrinsic region of the devices are ignored [14]. To accurately account for parasitic effects without over-burdening the circuit simulator, sub-circuit approaches employing lumped elements to represent device parasitics have gained popularity. These offer a good compromise between accuracy and speed [15].

In addition to parasitic effects, Non-Quasi-Static (NQS) effects become critical at higher frequencies, as the distributed RC effect inside the channel can no longer be ignored [16,17]. As the DC operating point in RF circuits moves from strong to weak inversion, an intrinsic MOSFET model that correctly models the turn-on of the transistor is essential. Furthermore, as oxide thickness aggressively scales to the direct-tunneling regime, the impact of gate leakage on device input impedance and minimum noise figure must be incorporated [18].

Substrate Noise Isolation

Digital CMOS devices are well known for the production of significant digital switching noise. CMOS digital circuits create short duration transients that generate both a continuous spectrum and a discrete spectrum (at multiples of the digital clock frequency). The clock harmonics are of particular importance in this marketplace because they may interact with the transmit/receive frequencies of communication elements.

Minimizing digital switching noise is a very difficult design issue [19]. Traditional noise-reduction methodologies include triple well (deep nwell), guard rings, careful attention to layout, and multiple voltage sources. In order to enhance the noise-isolation margin, noise-reduction strategies are frequently applied to both transmitters and receivers.

An interesting dilemma arises for higher frequency circuits (above 1GHz), where the advantages of many types of traditional noise isolation begin to fade. An example of this is provided in Figure 1, where simulation results are presented that compare triple well isolation and guard ring isolation as a function of frequency. Note that above 10GHz, the impact of traditional isolation methodologies is significantly reduced, and around 10GHz, methodologies such as guard rings offer equivalent benefits to triple well. In the <1GHz range, triple well on both the transmitters and receivers offers by far the best benefit.

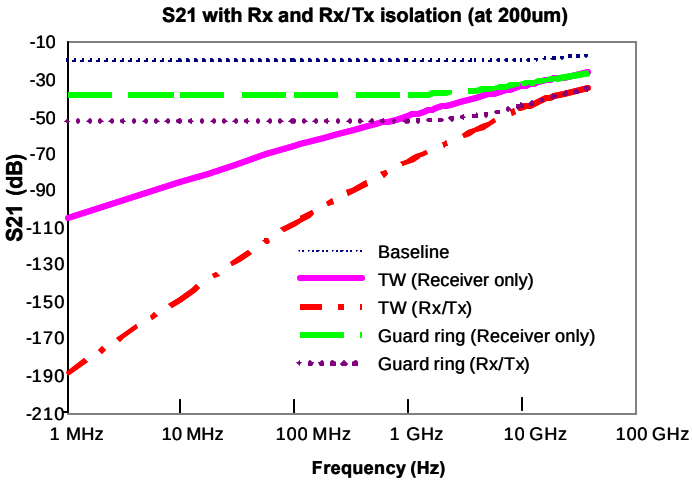


Figure 1: S_{21} noise isolation (in dB) comparing triple well to guard ring methodologies as a function of frequency

SiGe BJT DEVICES

SiGe epitaxial-base bipolar junction transistors are a key element in modern analog and high-frequency communications products. SiGe devices are the devices of choice for such applications as wireless Local Area Networks (LAN), 10G (with 40G on the horizon) synchronous optical networks (SONET) and in 1-2.5Gb/s Ethernet applications. SiGe devices also find application in the more traditional analog domain such as VCOs, mixers, power amplifiers, and Global Positioning Systems (GPS) devices.

The SiGe BJT is a richly researched device and only the high points will be covered here. Seminal review papers in 1995 and an update paper in 2001 provide an excellent summary of the field [20,21]. Table 3 summarizes key conference literature, indexed by company and performance level.

SiGe Performance Enhancement

The SiGe BJT provides performance enhancement in comparison with a conventional BJT device through three mechanisms.

The first enhancement arises from using the narrow SiGe bandgap to trade off against base and emitter implants. As can be seen in equation (1), decreasing the base bandgap energy (by adding Ge) permits an increase in base doping, which is desired, as it drops the base resistance, and a decrease in the emitter doping, which is also desired as it drops the emitter-base capacitance without seriously degrading . As can be seen from equations 2 and 3; lower R_b helps f_{max} and lower C_{eb} helps f_T. In addition (although not usually a limiter) the emitter transit time is inversely proportional to and thus benefits from the optimization.

$$\beta \propto \frac{N_{d(emitter)}}{N_{a(base)}} \times e^{\frac{E_{g(emitter)} - E_{g(base)}}{kT}} \quad (1)$$

$$f_T = \frac{1}{2\pi}\left[\frac{kT}{qI_c}\left(C_{eb}+C_{bc}\right)+\tau_b+\tau_e+\tau_{bc}\right]^{-1} \quad (2)$$

$$f_{\max} = \sqrt{\frac{f_T}{8\pi R_b C_{bc}}} \quad (3)$$

The second enhancement arises from tailoring the germanium profile (typically by ramping it across the rising edge of the base profile) in such a way as to accelerate electrons across the base and reduce the base transit time $\tau_b = W_B^2 / 2D_{nb}$.

Table 3: Comparison of key SiGe parameters as obtained from recent conference publications

Company	P. Author	FT	Fmax	Bvceo	Size	Alignment	Reference
Hitachi	K. Oda	124	174	2.3	0.2 x 1	SA	IEDM 2001
Conexant	M. Racanelli	170	160	2	0.15 x 10	SA	IEDM 2001
Infineon	J. Bock	106	145	2.3	0.18 x 2.8	SA	IEDM 2001
IHP	B. Heinemann	100	130	2.5	0.42 x 0.84		IEDM 2001
Hitachi	K. Washio	76	180	2.5	0.2 x 1	SA	IEDM 2000
ULSI	T. Hashimoto	73	61	2.6	0.15 x 6.15		IEDM 2000
Infineon	J. Bock	85	128	2.5	0.2 x 2.8	SA	IEDM 2000
Bell/Lucent	M. Carroll	58	102	3	0.28 x 0.84	SA	IEDM 2000
Infineon	J. Bock	52	65	2.7	0.2 x 0.28	SA	IEDM 1999
Hitachi	K. Washio	90	107	2	0.2 x 2	SA	IEDM 1999
IHP	K.E. Ehwald	55	90	2	0.8 x 2.5		IEDM 1999
Lucent	C.A. King	52	70	2	0.28 x 1.68	SA	IEDM 1999
IBM	G. Freeman	90	90	2.7	0.25 x 2.25	SA	IEDM 1999
Bell/Lucent	M. Carroll	45	35	4	0.28 x 1.68	SA	IEDM 1999
IHP	D. Knoll	65	90	2	1x1		IEDM 1998
IBM	Historic	90	90	2.7	A = 0.15 um^2		0.18 um
IBM	Historic	47	65	3.35	A = 0.3 um^2		0.25 um
IBM	Historic	47	65	3.35	A = 0.39 um^2		0.5 um

The third enhancement is an improvement in the Early voltage due to the use of a graded-Ge profile. The Early voltage is effectively a measure of how much the base profile can be depleted under reverse bias on the collector-base junction. Therefore, the Early voltage is a function of Ge-grading and reaches a maximum for a triangular Ge profile.

Self-Aligned SiGe BJT Devices

A typical SiGe BJT device incorporates a very thin SiGe layer wedged between the larger emitter and the substrate collector (see Figure 2). Presently, there are two common device configurations for modern SiGe BJT devices integrated into a BiCMOS process.

A quasi-aligned SiGe device (Figure 2, view "a") aligns the extrinsic base implant to the emitter poly edge. This means that "link" region (circled and a key contributor to R_b) is controlled by the interaction between two lithography layers (1 = the emitter cut and 2 = the emitter poly).

In contrast, the fully self-aligned device (Figure 2, view "b") uses a replacement emitter (usually called the emitter pedestal) and a spacer process to define the location of the extrinsic base implant. In this case, the extrinsic base is now "self-aligned" to the emitter as only one lithography operation (1 = emitter pedestal definition) is used to define the emitter cut and extrinsic base relationship.

The fully self-aligned device shows higher performance in the literature, but is somewhat more difficult to integrate due to the need to develop a replacement emitter process. As a consequence, the quasi-aligned process is frequently the more economical of the two.

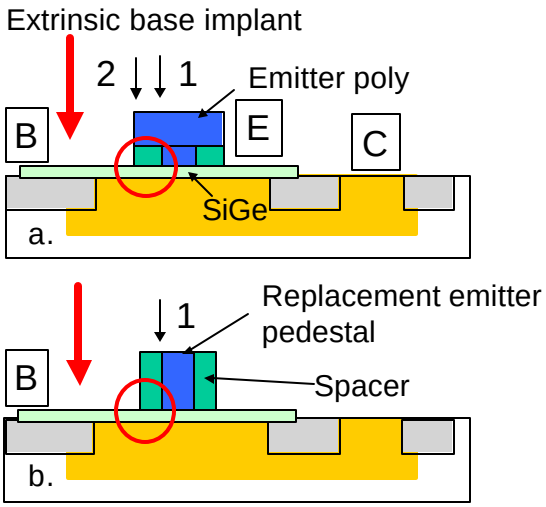


Figure 2: Comparison between quasi- and fully self-aligned SiGe bipolar transistor geometries

SiGe Versus CMOS

As the product marketplace moves to increasing integration of analog and mixed-signal elements with conventional digital CMOS, there are increasing demands on the process to economically integrate complex CMOS, BiCMOS, and bipolar process flows. With well-laid out CMOS devices showing f_T and f_{max} values in excess of 140GHz [12,22], a critical competitive question is "Why can't SiGe devices be replaced by CMOS?"

Table 4 attempts to answer this question by providing a high-level comparison between the key devices. CMOS devices offer the advantages of high f_T and f_{max} as well as superior linearity and lower voltage operation, due to lower threshold voltages (CMOS V_T in comparison to bipolar V_{BE}). BJT devices offer the advantages of excellent

noise performance and an improved transconductance (analog BJT in comparison to digitally-optimized CMOS). Also of interest in the comparison are density differences. BJT devices operating in low-noise amplifier applications occupy one-quarter to one- third the area of CMOS circuits of equivalent functionality. For the reverse example, CMOS devices operating in dense caches occupy one quarter to one third of the area of BJT circuits of equivalent functionality.

Note that noise is perhaps the major concern for CMOS RF design. The noise is due to the presence of interface states that introduce carrier trapping/de-trapping (1/f noise) and surface-roughness scattering (thermal noise). The coupling between the MOSFET channel and the gate also induces noise on the gate node at high frequencies.

Table 4: Comparison of CMOS with conventional and SiGe BJTs (summarized from Harame [1])

Parameter	CMOS	Si BJT	SiGe BJT
Ft	High	High	High
Fmax	High	High	High
Linearity	Best	Good	Better
Vbe (or VT) tracking	Poor	Good	Good
1/f noise	Poor	Good	Good
Broadband noise	Poor	Good	Good
Early voltage	Poor	OK	Good
transconductance	Poor	Good	Good

Designing Without BJT Devices

From a commercial CMOS perspective, the economical answer is to remove the BJT devices from the design. A SiGe BJT process adds between four and six masks to the conventional CMOS process, as well as a number of additional etches and thermal cycles that potentially damage the performance of the base CMOS devices. An ideal process would deliver SiGe performance using only CMOS.

Designing out the SiGe devices is an effort requiring both design and process contributions. From the design side, there is the requirement to design CMOS circuits that compensate for the poor noise performance. From the process side, there is the requirement to improve the analog performance of devices derived from a conventional digital process.

However, in the shorter term, the significant performance improvements offered by SiGe devices may validate the increased cost and complexity of integrating them into a full process flow.

PASSIVE ELEMENTS

A key difference between digital and mixed-signal processes is the presence of passive elements.

In the digital design world, performance is typically not determined by passive design elements. Capacitors are used as decoupling capacitors, and resistors are peripherally employed in IO-circuitry (in general, there are no intentionally fabricated inductors).

In strong contrast, in analog/mixed-signal design, performance is ultimately limited by the accuracy of the passive components in the technology [23,24,25,26,27]. In analog/mixed-signal design, passives (inductors, resistors, and capacitors) are used for a variety of active functions such as tuning, filtering, impedance matching, and gain control. Passives are key building blocks for circuits such as low offset voltage op-amps, analog frequency tuning circuits, switched capacitor circuits, filters, resonators, up-conversion and down-conversion mixers, VCOs, and D/A-A/D converters. The ability to accurately construct and model passives with Qs> 15-20 at frequencies >10G represents a key enabler for new circuits and products.

Inductors

Inductors are critical components in analog/mixed-signal design. Small-valued, precise, high-Q inductors are employed in circuits such as RF transceivers. Larger, lower-Q devices have functions such as impedance matching and gain control. Significant research has been done on monolithic integration of inductors, and in recent years there has been increasing use of inductors in state-of-the-art CMOS processes [28, 29, 30].

Spiral inductors in lengths can be fabricated with a conventional MOS process with negligible modifications to the design rules. A minimum of two metal layers is required, one to form the spiral and one to form the underpass. To minimize parasitic capacitance to the substrate, the top metal layer is the usual choice for the main spiral.

The most critical factor in inductor design is the optimization of the inductor Q at the design frequency. Q, or the "quality factor," is the ratio of the imaginary to the real part of the impedance (Q = Im(Z)/Re(Z)) and represents the ratio of the *useful magnetic stored energy* over the *average dissipation* for one cycle of the signal propagation. Note that determining the geometry and area required to deliver an optimized Q at the design frequency is not a straightforward process [31,32].

The most difficult factor in inductor process design is minimization of the impact of parasitic elements. Real inductors have parasitic resistance and capacitance. The parasitic resistance dissipates energy through ohmic loss, while the parasitic capacitance stores the unwanted energy. At high frequencies, the skin effect causes a non-

uniform current distribution in the metal segments, which introduces (among other things) a frequency-dependent contribution to the parasitic resistance. Finally, electromagnetic effects caused by the Faraday effect introduce parasitic currents (eddy currents) in the silicon as well, adding an additional frequency dependent term in the resistance [33].

Parasitic resistance is primarily driven by ohmic resistive losses in the thin patterned metal layers [34]. Parasitic resistance can be modulated both by design (trading off inductor area for inductor line width [35]) and by process (improving a Cu-damascene polish process to minimize dishing and thus permit wider metal lines).

Capacitive-induced loss is driven both by the Cox between the inductor and the substrate and by the lossy properties of the substrate. (At high frequencies the current flows through Cox and into the lossy substrate. The resulting dissipation adds a real component to the imaginary inductive impedance and degrades the Q.)

Minimizing this capacitance typically means separating the inductor as far as possible from the lossy silicon (usually by placing the inductor in the top metal layer). Recent advancements in low-k processes for digital CMOS also carry significant benefit (up to 4X improvement in Q for SiLK compared to conventional oxide ILD [35].)

Minimizing the substrate loss is more complex. As the frequency increases to where the skin depth is on the order of the substrate thickness, eddy currents in the substrate become a major loss mechanism. (This magnetically induced loss can be thought of as transformer action between a lossy primary and a lossy secondary [27].)

Mitigating eddy current loss can be quite difficult. There are a number of potential techniques including solid [33] and patterned ground shields [27], multilevel metalizations to build vertical solenoids [36], as well as minimizing doping levels under the inductor [33]. Note that since the eddy current loss is approximately proportional to the cube of the inductor diameter, strategies to minimize resistive parasitics by making large inductors (as is common in GaAs) are less effective in CMOS due to the more conductive Si substrates [27].

Capacitors

Analog/mixed-signal processes use four major types of capacitors. Polysilicon-insulator-polysilicon (PIP), metal-insulator-metal (MIM), lateral flux (finger), and MOS-style (depletion or accumulation).

Many older technologies have successfully used PIP capacitors. PIP capacitors do suffer from limited RF capability in the GHz range due to both the resistive losses in the plates and contacts, and to the parasitic capacitance between the passive element and the lossy silicon substrate [35]. Note also that the poly in PIP capacitors is typically implanted at higher doses than CMOS source-drain regions in order to minimize poly-depletion effects. This requires extra processing (and cost) because of additional lithography layers that need to be added to support the implants.

By far the most popular analog/mixed-signal capacitor is the metal-insulator-metal (MIM). MIM capacitors have the inherent advantage that they are metal (poly depletion and doping are non-issues) and, if implemented at the last metal layer, have the entire ILD stack between them and the substrate.

In recent years, the increasing interest in analog/mixed-signal commercial processes has led to implementation of MIM caps in commercial CMOS Cu-damascence processes [37,38,39]. The excellent linearity with voltage and temperature illustrates the popularity of the device as an analog element.

MIM devices are not without issues. Of special concern for today's processes is Cu metallurgy and its impact on yield and reliability. Also noteworthy is the choice of the inner layer dielectric. SiN is a popular choice due to common availability of the material in the traditional back-end process. However, note that low-temperature deposited SiN is known to show higher relaxation recovery voltages than oxide [40]. PECVD SiN displays significant sensitivity to operation frequency, bias voltage, and temperature when compared to oxide [41]. SiN also displays frequency dependent shifts that are consistent with bulk-nitride-traps [42] located within a tunneling distance of the nitride metal interface.

One of the restrictions with MIM devices is that process technologies do not scale the vertical spacing in the back end nearly as fast as the lateral spacing. The reason is that digital circuit designs cannot tolerate large increases in the wiring capacitance from generation to generation. Lateral flux (finger) capacitors solve this problem by using the lateral capacitance (between the metal lines) rather than the vertical capacitance (between the different ILD layers). As a consequence, the capacitance is under design control and scales more effectively with the technology [27].

Another of the limitations of the MIM device is the thickness of the insulator region. In contrast, MOS

Other brands and names are the property of their respective owners.

devices can take advantage of thin gate oxide processes to achieve high capacitance per unit area. However, since one of the contacts is formed in silicon, the series resistance of a MOS capacitor is quite large. In addition, the very high gate leakage currents of modern scaled oxides (180 node and beyond, or <30A electrical) make gate-oxide-based MOS devices excessively leaky for conventional applications.

RESISTORS

Precision polysilicon and metal thin film resistors are key passive elements in analog circuits. The simultaneous presence of both poly and metal resistors can add value in a process, because the metal resistors are at the top of the stack and the poly resistors at the bottom. Two widely separated locations allow designers to chose a resistor that minimizes parasitics for their particular circuit. Also, the presence of a front-end resistor may enable in-line or early learning electrical evaluation on key circuit elements.

NAME	TYPE	Rho (ohm/sq)	VCR (ppm/V)	TCR (ppm/C)
Foundary A (N+)	N+ poly	126	-550	46
Foundary A (P+)	P+ poly	360	-56	-187
Foundary A HR	HR poly	1000	-70	-1250
Foundary B (N+)	N+ poly	77	210	46
Foundary B (P+)	P+ poly	258	519	148
Stuber (PS)	Polycide	12	320	440
Stuber (N+)	N+ poly	145	150	640
Stuber (P+) [45]	P+ poly	225	100	1440
Jeng #C (P+)	P+ poly	274		-96
Jeng #D (P+)	P+ poly	306		-285
Jeng #F (P+) [46]	P+ poly	244		-40

Table 5: Comparison of various poly resistors as reported by commercial sources and reviewed in the literature

Polysilicon Resistors

Polysilicon resistors exist in both silicided and unsilicided versions. Since the resistance of polysilicon-silicided (polycide) resistors tends to be quite low (5-15 ohms/sq), and the voltage coefficient tends to be quite high (100-600 ppm/V) there is a strong tendency to use the unsilicided (or silicide-blocked) resistors.

In a typical silicide-blocked resistor, the center of the device is silicide-blocked and the endcaps are left open. The endcaps either receive the conventional silicide processing for a contact pad, or receive optimized processing specific to the resistor application [43].

The silicide-blocking layer is usually an oxide or nitride and is frequently chosen to leverage a pre-existing layer elsewhere in the process. Existence of a silicide-blocking layer also enables devices such as silicide-blocked diffusion resistors (see Figure 3) and silicide-blocked MOS devices [44].

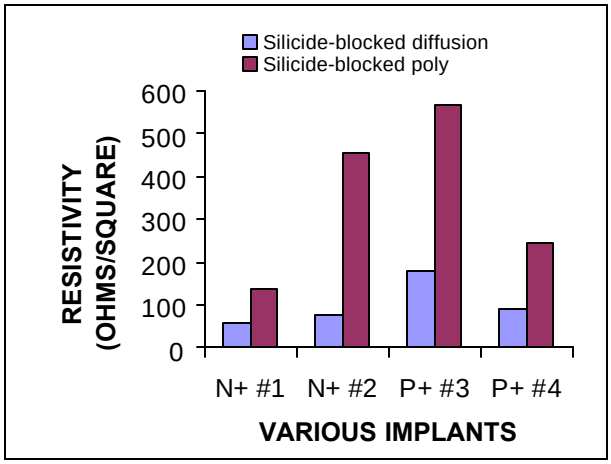


Figure 3: Resistivity for silicide-blocked diffusion and polysilicon resistors

Polysilicon resistors are usually placed on a field. In technologies with thin field oxides (such as LOCOS), there is significant electrical interaction through the field oxide and parasitic capacitance as well as depletion of the bottom of the resistor, which produces a voltage-dependent resistance change. All these must be considered in the resistor design [47]. Such effects are significantly reduced with the thicker oxides (3000-6000A), characteristic of STI processes, and they are also significantly reduced with SOI processing [48].

The sheet resistance, as well as the thermal and voltage coefficients of silicide-blocked polysilicon resistors, are very process-dependent. Implant conditions, grain boundary size, thermal activation, and end-cap silicide quality can all impact the key polysilicon resistor parameters. As a consequence, reported values for the major resistor parameters vary widely. Table 5 provides a cross-section of industry values, and Figure 3 compares values within a single Intel process for both silicide-blocked diffusion and poly devices.

Metal Film Resistors

Metal thin film resistors can be built at any of the traditional metal layers. In addition, a metal thin film resistor can be built as a by-product of the MIM capacitor

process. TaN is frequently used as well, due to its ready availability in a Cu-damascene process as a Cu-diffusion barrier. TaN is also interesting to the process designer as it exhibits a TCR-versus-resistivity relationship that ranges from roughly 500 ppm/C at 50 ohms/sq. to roughly (-)500 ppm/C at 400 ohms/sq and is attributed to the transition from metallic conduction (positive TCR) to hopping conduction (negative TCR). Zero TCR is ~250 ohms/sq. [39]. (A similar effect is also seen in W-silicide resistors, with a transition point at ~40 ohms/sq. [49].)

CONCLUSION

Analog/mixed-signal/RF continues to be a challenge for digital CMOS designers and manufacturers. Conflicting scaling methodologies, complex measurement and modeling support requirements, a multiplicity of interacting features, and increasingly complex process integration issues are the challenges to overcome to support the next generation of product designs.

ACKNOWLEDGMENTS

We acknowledge the significant contributions of Ian Young, Paul Packan, Chris Kardas, Doug Barlage, Perry Heedley, Rafael Rios, Niraj Anand, Issy Kipnis, Mitch Denham, Fred Buckley, Kim-Ahn Huynh, Steve Dreyer, Jian Gong, and Atul Shah.

REFERENCES

[1] D. Harame, "High Performance BiCMOS Process Integration: Trends, Issues and Future Directions," *IEEE BTCM 2.1* pp. 36-43.

[2] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 1999 ed. Austin, TX., International SEMATECH.

[3] S. Song, J.H. Yi, W.S. Kim, J.S. Lee, K. Fujihara, H.K. Kang, J. T. Moon, and M.Y. Lee, "CMOS device scaling beyond 100nm," *IEDM 2000, 10.4.1.*

[4] M. Bohr and Y. El-Mansy, "Technology for Advanced High-Performance Microprocessors, " *IEEE Trans. Electron Dev.*, vol. 45, no. 3, March 1998, pp. 620-625.

[5] S. Thompson, P. Packan, M. Bohr, "MOS Scaling: Transistor Challenges for the 21st Century," *Intel Technology Journal*, July 1998.

[6] S. Thompson, et al., "An Enhanced 130nm Generation Logic Technology Featuring 60nm Transistors Optimized for High Performance and Low Power at 0.7-1.4 V," *IEDM 2001*, pp. 257-260.

[7] P.A. Stolk, H.P. Tuinhout, R. Duffy, E. Augendre, L.P. Bellefroid, M.J.B. Bolt, J. Croon, C.J.J. Dachs, F.R.J. Huisman, A.J. Moonen, Y.V. Ponomarev, R.F.M. Roes, M. Da Rold, E. Sevinck, K.N. Sreerambhatla, R. Surdeanu, R.M.D.A. Velghe, M. Vertregt, M.N. Webster, N.K.J. van Winkelhoff, A.G.A. Zegers-Van Duijnhoven, "CMOS Device Optimization for Mixed Signal Technologies," Electron Devices Meeting, 2001, *IEDM Technical Digest International*, 2001, pp. 10.2.1 -10.2.4.

[8] A. Chatterjee, K. Vasanth, D. Grider, M. Nandakumar, B. Pollack, R. Aggarwal, M. Rodder, H. Shichijo, "Transistor design issues in integrating analog functions with high performance digital," VLSI Technology, 1999, *Digest of Technical Papers*, 1999 Symposium, 1999 pp. 147 -148.

[9] P. Woerlee, M. Knitel, R. Langevelde, D. Klaassen, L. Tiemeijer, A. Scholten, and D. Zegers-van Duijnhoven, "RF-CMOS Performance Trends," *IEEE Trans. on Elec. Devices*, Vol. 48, No. 8, August 2001.

[10] A. Stolk, H.P. Tuinhout, R. Duffy, E. Augendre, L.P. Bellefroid, M.J.B. Bolt, J. Croon, C.J.J. Dachs, F.R.J. Huisman, A.J. Moonen, Y.V. Ponomarev, R.F.M. Roes, M. Da Rold, E. Sevinck, K.N. Sreerambhatla, R. Surdeanu, R.M.D.A. Velghe, M. Vertregt, M.N. Webster, N.K.J. van Winkelhoff, A.G.A. Zegers-Van Duijnhoven, "CMOS Device Optimization for Mixed Signal Technologies," Electron Devices Meeting, 2001, *IEDM Technical Digest International,* 2001, pp. 10.2.1-10.2.4.

[11] E. Vandamme, D. Schreurs, and D. Dinther, "Improved Three-step De-embedding Method to Accurately Account for the Influence of Pad Parasitics in Silicon On-Wafer RF Test Structures," *IEEE Trans. On Electron Devices*, Vol. 48, No. 4, April 2001.

[12] L.F. Tiemeijer, H.M. Boots, R.J. Havens, A.J. Scholten, P.H. W. de Vreede, P.H. Woerlee, A.Heiringa, and D. B. M. Klaassen, "A record high 150 GHz fmax realized at 0.18um gate length in an industrial RF-CMOS technology," Electron Devices Meeting, 2001, *IEDM Technical Digest International*, 2001, pp. 10.4.1-10.4.4.

[13] T. H. Lee, *The Design of CMOS RF Integrated Circuits*, Cambridge Univ. Press, chapter 10, 1998.

[14] William Liu, MOSFET Models for SPICE Simulation Including BSIM3 and BSIM4, John Wiley & Sons, Inc., pp. 316-329, 2001.

[15] Christian Enz, "An MOS Transistor Model for RF IC Design Valid in All Regions of Operation," *IEEE*

Trans. Microwave Theory & Tech., Vol. 50, No.1, pp. 342-359, 2002.

[16] W. Liu et al., "A CAD-Compatible Non-quasistatic MOSFET Model," *IEDM Tech. Digest*, pp. 151-154, 1996.

[17] E. Gondro et al., "When Do We Need Non-Quasistatic CMOS RF Models?", *IEEE Custom IC Conference*, pp. 377-380, 2001.

[18] R. van Langevelde et al., "Gate Current: Modeling, DL Extraction and Impact on RF Performance," *IEDM Tech. Digest*, pp. 289-292, 2001.

[19] R. Frye, "Integration and electrical isolation in CMOS mixed-signal wireless chips," in Proceedings of the *IEEE*, Vol. 89, No. 4, April 2001.

[20] D.L. Harame, J.H. Comfort, J.D. Cressler, E.F. Crabbe, J.Y-C Sun, B.S. Meyerson, and T. Tice, "Si/SiGe Epitaxial-Base Transistors – Part I: Materials, Physics, and Circuits," *IEEE Transactions on Electron Devices*, Vol. 42, No. 3, March 1995 pp. 455-468, and Si/SiGe Epitaxial-Base Transistors – Part II: Process Integration and Analog Applications," *IEEE Transactions on Electron Devices*, Vol. 42, No. 3, March 1995, pp. 469-482.

[21] D. Harame, D. Ahlgren, D. Coolbaugh, J. Dunn, G. Freeman, J. Gillis, R. Groves, G. Hendersen, R. Johnson, A. Joseph, S. Subbannna, A. Victor, K. Watson, C. Webster, and P. Zampardi, "Current Status and Future Trends of SiGe BiCMOS Technology," *IEEE Transactions on Electron Devices*, Vol. 48, No. 11, November 2001, pp. 2575-2594.

[22] Y. Momiyama, T. Hirose, H. Kurata, K. Goto, Y. Watanabe, and T. Sugii, "A 140 GHz ft and 60 GHz fmax DTMOS integrated with high-performance SOI logic technology," Electron Devices Meeting, 2000, *IEDM Technical Digest International*, 2000, pp. 451-454 (Fujitsu).

[23] D. J. Allstot and W.C. Black, Jr., "Technology Design considerations for monolithic MOS switched-capacitor filtering systems," in *Proceedings IEEE*, Vol. 71, pp. 967-986, August 1983.

[24] J. J. Wikner and N. Tan, "Influence of circuit imperfections on the performance of DACs," *Analog Integrated Circuits Signal Processing*, Vol. 18, pp. 7-20, 1999.

[25] J. L. McCreary, "Matching properties and voltage and temperature dependence of MOS capacitors," *IEEE J. Solid-State Circuits*, Vol. SC-16, pp. 608-616, June 1981.

[26] R.K. Ulrich et al., "Getting Aggressive with Passive Devices," *IEEE Circuits and Devices*, Sept. 2000, pp. 17-25.

[27] Thomas H. Lee and S. Simon Wong, "CMOS RF Integrated Circuits at 5GHz and Beyond," in *Proceedings of the IEEE*, Vol. 88, No. 10, October 2000.

[28] K. B. Ashby, W.C. Finley, J.J. Bastek, and S. Moinian, "High Q inductors for wireless applications in a complementary silicon bipolar process," in *Proceedings Bipolar/CMOS Circuits and Technology Meeting*, 1994, pp. 179-182.

[29] J.N. Burghartz, D. C. Edelstein, K. A. Jenkins, C. Jahnes, C. Uzoh, E.J. O'Sullivan, K.K. Chan, M. Soyuer, P. Roper, and S. Cordes, "Monolithic spiral inductors fabricated using a VLSI Cu-damascene interconnect technology and low loss substrates," in *Technical Digest International Electron Devices Meeting*, 1996, pp. 99-102.

[30] J.N. Burghartz, M. Soyuer, and K.A. Jenkins, "Microwave Inductors and capacitors in a standard multi-level interconnect silicon technology," *IEEE Trans. Microwave Theory*, Tech. Vol. 44, No. 1, pp. 100-104, 1996.

[31] J.R. Long and M.A. Copeland, "Modeling, characterization and design of monolithic inductors for silicon RF IC's," in *Proceedings Custom Integrated Circuits*, Conference 1996, pp. 155-158.

[32] C.P. Yue, C. Ryu, J. Lau, T.H. Lee, and S. S. Wong, "A physical model for planar spiral inductors on silicon," in *Technical Digest International Electron Devices Meeting*, 1996, pp. 155-158.

[33] J. N. Burghartz, M. Soyuer, K. Jenkins, M. Kies, M. Dolan, K. J. Stein, J. Malinowski and D. Harame, "Integrated RF Components in a SiGe Bipolar Technology," *IEEE Journal of Solid State Circuits*, Vol. 32, No. 9, September 1997.

[34] N. M. Nguyen and R.G. Meyer, "SiIC-compatible inductors and LC passive filters," *IEEE J. Solid-State Circuits*, Vol. 25, No. 4, pp. 1028-1031, 1990.

[35] Snezena Jeni, Stefaan Decoutere, Stefann Van Huylenbroeck, Guido Vanhorebeek and Bar Nauwelaers, "High-Q Inductors and Capacitors on Si substrate," 2001 Topical Meeting on Silicon Monolithic Integrated Circuits in *RF Systems, 2001 Digest of Papers*, pp. 64-70.

[36] J.N. Burghartz, M. Soyuer, and K.A. Jenkins, "Integrated RF and microwave components in

BiCMOS technology," *IEEE Trans. Electron Devices*, Vol. 43, No. 9, pp. 1559-1570, 1996.

[37] M. Armacost, A. Augustin, P. Felsner, Y. Feng, G. Friese, J. Heidenreich, G. Hueckel, O. Prigge, and K. Stein, "A high reliability metal insulator metal capacitor for 0.18um copper technology," *IEDM 2000* paper 7.4.1.

[38] R. Liu, et al., "Single Mask Metal-Insulator-Metal (MIM) Capacitor with Copper Damascene Metallization for Sub-0.18um Mixed Mode Signal and System-on-a-Chip (SoC) Applications," in *Proceedings IITC*, 111 (2000).

[39] P. Zurcher, P. Alluri, P. Chu, A. Duvallet, C. Happ, R. Henderson, J. Mendonca, M. Kim, M. Petras, M. Raymond, T. Remmel, D. Roberts, B. Steimle, J. Stipanuk, S. Straub, T. Sparks, M. Tarabbia, H. Thibieroz, and M. Miller, "Integration of Thin Film MIM Capacitors and Resistors into Copper Metallization based RF-CMOS and Bi-CMOS Technologies," Electron Devices Meeting, 2000, *IEDM Technical Digest International*, 2000, pp. 153 - 156.

[40] J. Fattaruso, et al., *IEEE Journal of Solid State Circuits*, Vol. 25, No. 12, (1990).

[41] J. Babcock, S. Balster, A. Pinto, C. Dirnecker, P. Steinmann, R. Jumpertz and B. El-Kareh, "Analog Characteristics of Metal-Insulator-Metal Capacitors using PECVD Nitride Dielectrics," *IEEE Electron Device Letters*, Vol. 22, No. 5, May 2001, pp. 230-232.

[42] W. S. Lau, "The identification and suppression of defects responsible for electrical hysteresis in metal-nitride-silicon capacitors," *Japan J. Appl, Phys. Lett*ers, Vol. 29, No. 5, pp. L690-693, 1990.

[43] Wen-Chau Lin, Kong-Beng Thei, Hung-Ming Chuang, Ken-Wei Lin, Chin-Chuan Cheng, Yen-Shih Ho, Chi-Wen Su, Shyh-Chyi Wong, Chih-Hsien Lin and Carlos Diaz, "Characterization of polysilicon resistors in sub-0.25um CMOS ULSI applications," *IEEE Electron Device Letters*, Vol. 22, No. 7, July 2001.

[44] A. Salman, R. Gauthier, W. Stadler, K. Esmark, M. Muhammad, C.Putnam, D. Ioannou, **"**Characterization and investigation of the interaction between hot electron and electrostatic discharge stresses using NMOS devices in 0.13 /spl mu/m CMOS technology," Reliability Physics Symposium, 2001, in *Proceedings 39th Annual 2001 IEEE International*, 2001, pp. 219 - 225.

[45] M. Stuber, M. Megahed, J. Lee, T. Kobayashi, H. Domyo, "SOI CMOS with higher performance passive components for analog, RF and mixed signal design," in *Proceedings 1998 IEEE Int. SOI conference*, October 1998, pp. 99-100.

[46] S.J. Jeng, D. C. Ahlgren, G. D. Berg, B. Ebersman, G. Freeman, D.R. Greenberg, J. Malinowski, D. Nguyen-Ngoc, K. T. Schonenberg, K. J. Stein, D. Colavito, M. Longstreet, P. Ronsheim, S. Subbanna, D. L. Harame, "Impact of Extrinsic Base Process on NPN HBT Performance and Polysilicon Resistor in Integrated SiGe HBTs," *IEEE BCTM 12.1* 1997, pp. 187-190.

[47] Yannis Tsividis, *Mixed Analog-Digital VLSI Devices and Technologies*, McGraw-Hill, 1996, pp. 163-169.

[48] M. Stuber, M. Megahed, J. Lee, T. Kobayashi, H. Domyo, "SOI CMOS with higher performance passive components for analog, RF and mixed signal design," in *Proceedings 1998 IEEE International SOI conference*, Oct. 1998, pp. 99-100.

[49] C.S. Pai, M. K. Bude, M. Frei, S. Rogers, D. Jacobson, S. Merchant, F. Hui, R. Liu and R. Gregor, "Integrated W-silicide metal resistor for advanced CMOS technologies," Interconnect Technology Conference, 2001, in *Proceedings of the IEEE 2001 International*, 2001, pp. 216 -218.

AUTHORS' BIOGRAPHIES

Kelin Kuhn received an M.S. and a Ph.D. E.E. degree from Stanford in 1985. She works in technology development at Intel PTD. Prior to joining Intel, she was an Electrical Engineering professor at the University of Washington and has written technical papers and a textbook in the area of photonics and microelectronics. Her e-mail is kelin.ptd.kuhn@intel.com.

Shahriar Ahmed received his Ph.D. E.E. degree from Rice University in 1984. He is a senior RF/mixed-signal device engineer in Intel PTD. He has worked for Intel for 17 years and his expertise is in bipolar, biCMOS, and CMOS devices. His e-mail is shahriar.ahmed@intel.com.

Peter Vandervoorn received his B.S. degree in 1992 and his Ph.D. E.E. degree in 1998 from Cornell University. He is a Senior Process Integration Engineer in PTD. He has been with Intel PTD for four years and his previous position was in integration for Intel's 0.13um CMOS process. His e-mail is peter.j.vandervoorn@intel.com.

Anand Murthy received his Ph.D from University of Southern California and joined Intel in 1995. His current focus includes thin film deposition for novel transistor research. His e-mail address is anand.murthy@intel.com

Kartik Raol received his B.S. E.E. degree from the University of Texas at Austin in 1984 and his M.S. E.E. degree from the University of Cincinnati in 1987. He is a Staff Engineer with the Compact Device Modeling Group in Technology CAD (LTD). Prior to joining Intel in 1995 he was a Principal Engineer (Technology CAD) at Digital Equipment Corporation, Hudson, MA. His primary interests are in the areas of interconnect and device modeling/simulation and statistical circuit design techniques. His e-mail is kartik.raol@intel.com.

Borna Obradovic received a Ph.D in Electrical Engineering from the University of Texas at Austin in 1999. He currently works for the TCAD dept. at Intel Corp., specializing in process and device simulation/ modeling. His e-mail is borna.obradovic@intel.com.

Wei-Kai Shih joined Intel TCAD in 1997 after obtaining his Ph.D. degree in Electrical Engineering from the University of Texas at Austin. Since then he has been involved in various reliability and compact modeling projects. He played a key role in developing Intel's chip thermal simulation tool and made significant contributions to the development of industry-leading compact models for advanced MOS transistors. His current interests are in modeling the high-frequency behavior of MOS devices and studying the impact of quasi-ballistic carrier transport in sub-100nm transistors. His e-mail is wei-kai.shih@intel.com.

Iwen Chao received his Ph.D. degree in Electrical Engineering from the University of California, Davis in 1992. He joined Intel FCD in 1995 and worked on the first Intel StrataFlash® memory project. Since 2000, he has been the silicon technology manager of the Method and Technology Group in Intel ICG, primarily working on silicon technology definition for mix-signal circuits in high speed Ethernet and optical products. His e-mail is iwen.chao@intel.com.

Ian Post received his Ph.D. E.E. degree from the University of South Hampton in 1992. He is a senior device engineer in Intel PTD. He has worked for Intel for seven years and his present expertise is in bipolar, biCMOS, and CMOS devices. Prior to joining PTD he worked for Fab 7 in flash integration. His e-mail is ian.r.post@intel.com.

Stephen Chambers received his M.S. E.E. from California Institute of Technology in 1979. He is a Senior Process Integration Engineer in PTD. He has been with Intel for 23 years and his interests are in advanced metal systems and interconnect technology. His e-mail is stephen.chambers@intel.com

Copyright © Intel Corporation 2002. This publication was downloaded from http://developer.intel.com/.

Legal notices at http://www.intel.com/sites/corporate/tradmarx.htm.

Transistor Elements for 30nm Physical Gate Lengths and Beyond

Brian Doyle, Technology and Manufacturing Group, Intel Corporation
Reza Arghavani, Technology and Manufacturing Group, Intel Corporation
Doug Barlage, Technology and Manufacturing Group, Intel Corporation
Suman Datta, Technology and Manufacturing Group, Intel Corporation
Mark Doczy, Technology and Manufacturing Group, Intel Corporation
Jack Kavalieros, Technology and Manufacturing Group, Intel Corporation
Anand Murthy, Technology and Manufacturing Group, Intel Corporation
Robert Chau, Technology and Manufacturing Group, Intel Corporation

Index words: transistor, scaling, SOI, DST, fully-depleted, high-k, double-gate

ABSTRACT

We have fabricated conventional planar transistors of various gate lengths down to as small as 10nm polysilicon gate lengths, in order to examine transistor scaling. At 30nm gate lengths, the devices show excellent device characteristics, indicating that this node can be met with conventional transistor design. At lower gate lengths of 20 and 15nm, the devices still maintain excellent device characteristics and follow traditional scaling with respect to gate delay and energy delay, although off-state leakage and gate leakage do increase. At 10nm gate lengths, the transistors continue to function as MOS devices, but they are limited by off-state leakage.

One feasible method of significantly improving off-state leakage is through reducing the sub-threshold gradient. We show that *Depleted Substrate Transistors (DST)*, a broad category of devices that include single- and double-gate transistors, whose active channel region stays fully depleted during operation, can achieve near-ideal sub-threshold gradients and a reduction in off-state leakage of at least two orders of magnitude over bulk transistors. We believe that DST architecture will adequately address transistor scaling needs down to 10nm gate lengths.

In addition to DST device architecture, new electronic materials and modules will be needed to maintain high performance and low-parasitic leakages. As an example, to alleviate increasing gate leakage, changes in the gate stack are necessary. Replacement of SiO_2, the workhorse of the industry for over 30 years, with a high-K dielectric will be required. Other changes will include use of raised source/drain, metal gate electrodes and channel engineering.

INTRODUCTION

Moore's Law, formulated in the 1960s, states that the transistor count on an integrated circuit chip doubles every 18 months and has been the driving force behind the phenomenal growth of the semiconductor industry. This same law, which is the basis for the International Technology Roadmap for Silicon, guides the industry with respect to the features of future generations [1].

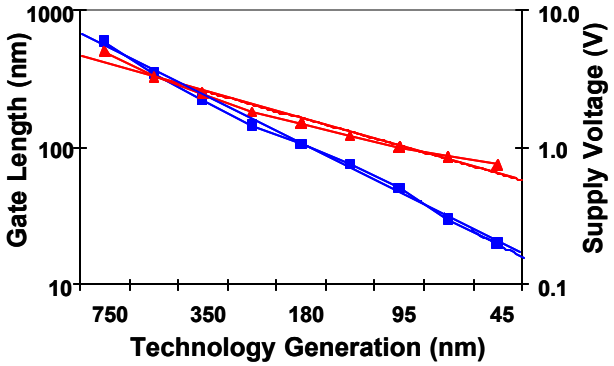


Figure 1: Gate length and power-supply voltage vs technology node

One of the most important consequences of scaling resulting from Moore's law is transistor gate length scaling. Figure 1 shows the gate lengths and power-supply voltages as a function of technology generation. Historically, the power supply scales at 0.85x/generation, while the gate length scales at 0.65x/generation. The gate length has been scaling and is expected to continue to scale at considerably less than half the lithography pitch for future generations. With respect to the power supply,

the voltage is expected to drop below one volt shortly, and continue to decrease. It is thus of considerable interest to study the implications of gate length and power-supply scaling for transistor design and architecture.

In this paper, we examine gate length scaling on bulk MOS devices. Although we concentrate mostly on n-MOS devices, the same results and conclusions are true for p-MOS transistors as well. Using special lithographic techniques, we show that it is possible to shrink gate lengths as small as 10nm and still maintain meaningful transistor functionality. The device characteristics of transistors at these small gate lengths and associated scaling issues are discussed in detail. Possible solutions to enable continued scaling are then proposed.

GATE LENGTH SCALING

Polysilicon Patterning

In order to examine the consequences of gate length scaling down to 10nm, a methodology for patterning polysilicon at these extreme dimensions, called *Spacer Gate* (SG) [2], has been used.

This approach has recently been shown to be capable of generating line widths down to 6.5nm [3]. The SG process steps are outlined in Figure 2.

After poly gate electrode deposition, an oxide layer (100nm) is deposited and patterned so that the edge of the oxide blocks is aligned to the edge of the gates to be patterned. A nitride film is deposited on the wafer, whose thickness determines the dimension of the gate to be printed. The nitride is now RIE-etched, leaving a nitride spacer on the oxide block sidewalls. The oxide block is now removed, and the polysilicon is then etched, leaving polysilicon lines whose dimensions are controlled simply by the thickness of the nitride film deposited.

There are several inherent advantages to the SG approach over conventional lithography:

> The dimensions of the lines being printed depend only on the thickness of the nitride layer deposited.

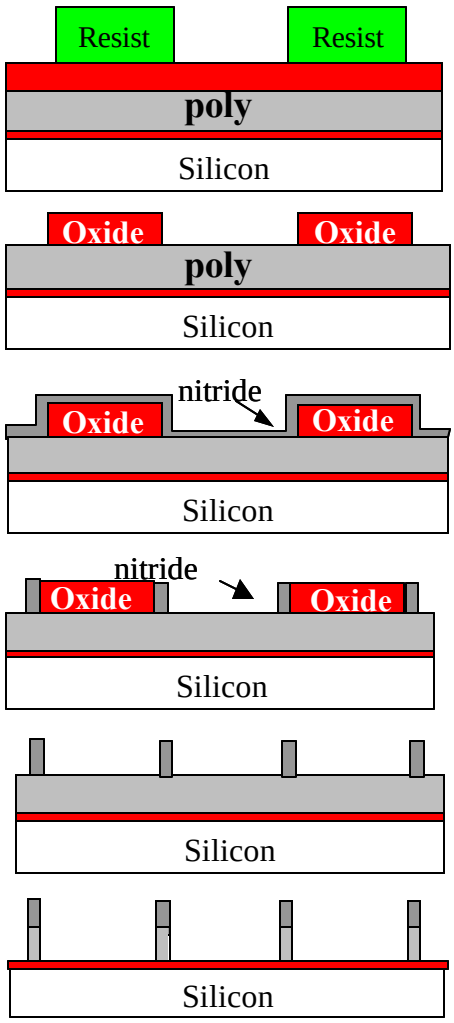


Figure 2: Spacer Gate flow

Since every oxide block produces two poly lines (see Figure 2), the pitch needed for the SG is half that of conventional lithography, and hence an n-2 generation lithography tool can be used to print nth-generation poly lines.

Critical Dimension (CD) control from SG would be expected to be superior to conventional lithography, since it depends simply on the deposition and etch of a thin film in contrast to the many factors that come into play in conventional lithography.

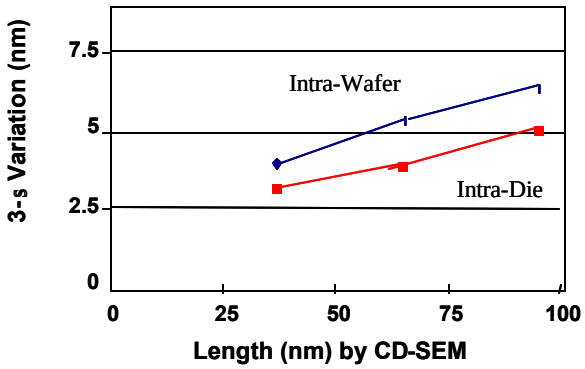


Figure 3: Gate length vs 3-s Electrical CD variation for structures fabricated by Spacer Gate techniques

In order to test for CD control with SG, Electrical CD (ECD) structures were measured (Figure 3). For structures measuring 95nm, 65nm, and 38nm, the intra-wafer 3- CD control measured 6.3, 5.3, and 3.9nm respectively. The intra-die values were even tighter, at 4.9, 3.8, and 3.1nm respectively, meeting the 10% 3- CD variation targets of the silicon technology roadmap [1].This SG approach enables the fabrication of polysilicon lines down to 10nm using 248nm lithography. Figure 4 shows an example of the methodology. The top-down Scanning Electron Microscope (SEM) micrograph of a polysilicon line, fabricated using a nitride film whose thickness was 30nm, resulted in poly lines with a width of 30nm. It can also be seen that the lines are extremely straight, showing very little line-edge roughness.

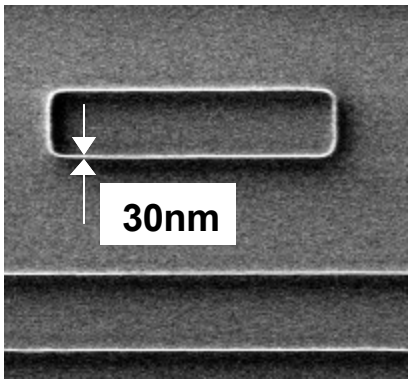


Figure 4: Top-down SEM of poly lines printed using the Spacer Gate technique

This technique was used for the fabrication of n-MOS transistors to explore transistor scaling. An example of a 15nm device fabricated using this technique is shown in Figure 5 .

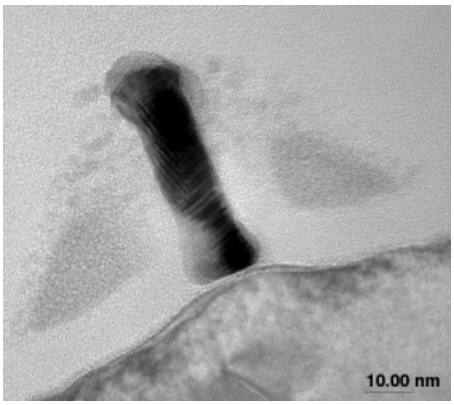


Figure 5: TEM cross-section of a 15nm transistor

For the devices discussed in this paper, the physical gate oxide was aggressively scaled to sub-1.0nm in order to achieve high drive currents and controllable short channel effects. Figure 6 (insert) shows a TEM cross-section of the sub-1.0nm gate oxide. Because of difficulties in measuring GV in the presence of high gate leakage, a transmission line model was used [4]. Figure 6 also shows the inversion GV characteristics of the resulting gate stack. An inversion capacitance exceeding 1.9 F/cm^2 was achieved for both p- and n-MOS.

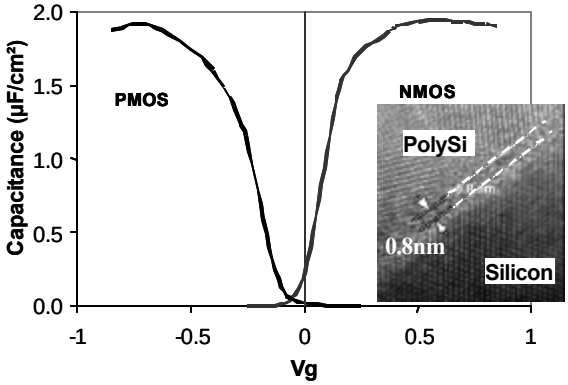


Figure 6: Inversion C-V capacitance

In order to control short channel effects and achieve sufficiently low external resistance and overlap capacitance, retrograded wells, aggressively scaled S/D and S/D extensions, and thermal anneal temperatures below 1000°C were used. To minimize the poly depletion effect with scaled junctions, the polysilicon gate thickness was scaled to below 100nm. The silicide was also scaled with junction scaling.

Figures 7 and 8 show the I-V characteristics of a 30nm device [5]. It can be seen that this transistor shows excellent I_{on}-I_{off} performance with I_{on}=570 A/ m for n-MOS and 285 A/ m for p-MOS with I_{off} at or below 100nA/ m at a scaled-down voltage of Vcc=0.85V.

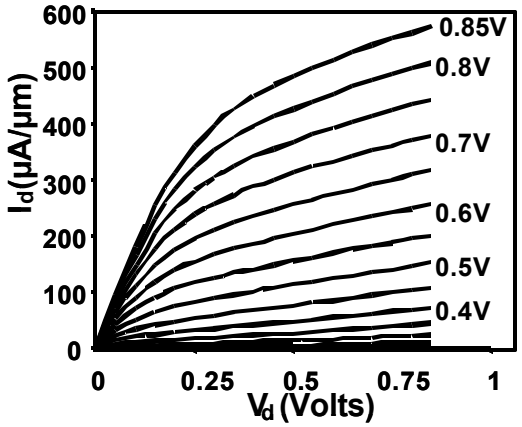


Figure 7: MOSFET I_d-V_d curves for the 30nm n-MOS

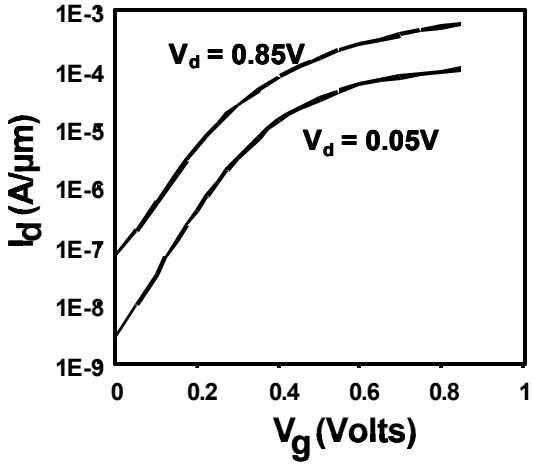


Figure 8: MOSFET sub-threshold I_d-V_g curves for the 30nm n-MOS device

Going to even shorter channel lengths, Figures 9 and 10 show the I-V characteristics of an n-MOS device with a 15nm physical gate length.

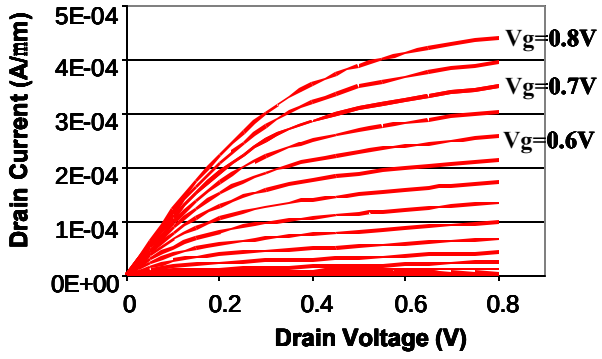


Figure 9: I_d-V_d characteristics for a n-MOS transistor with a physical gate length of 15nm

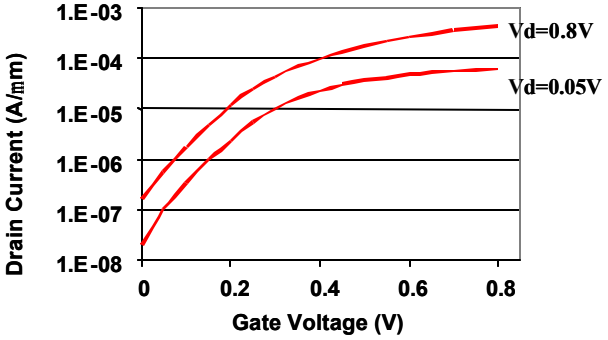


Figure 10: Id-V_g characteristics for the same 15 nm gate length device as in Figure 9

Figure 9 shows the I_d-V_d characteristics for different applied gate voltages, while Figure 10 shows the I_d-V_g at V_d=0.05V and V_d=0.8V. It can be seen that the device has I_{off} of 180nA/ m at V_{cc}=0.8V, a sub-threshold gradient of 95mV/decade at V_{cc}=0.85V, and a Drain Induced Barrier Lowering (DIBL – the gate voltage difference for Id=1E-6 A/ìm between Vd=0.05V and Vd=0.8V) effect of about 90mV/V. These results suggest that the device has a controllable short channel effect. The drive current for this device is 443 A/ m at V_{cc}=0.8V, as can be seen from the Id-Vd characteristics of Figure 9.

Moving to smaller dimensions, Figure 11 shows a cross-sectional TEM of a transistor with poly gate length measuring only 10nm. At these dimensions, even the slight recessing of the source-drain region becomes greatly magnified and tends to thicken up the gate oxide. Another aspect of the device is that the height to width of the transistor is approximately scaled, the height being approximately 50nm for this 10nm L_g transistor.

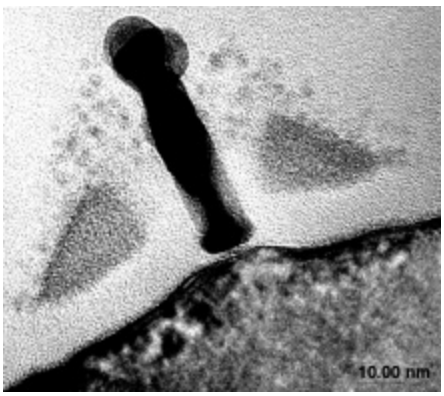


Figure 11: TEM cross-section of a 10nm transistor

To get a perspective of the magnitude of the scaling to get to 10nm, Figure 12 compares TEM's from the 0.18 micron technology node with the 10nm transistor (circled in this figure) on the same scale. It can be seen that the transistor is barely visible at this magnification, and that to

get these dimensions, not only the width, but the height is also scaled.

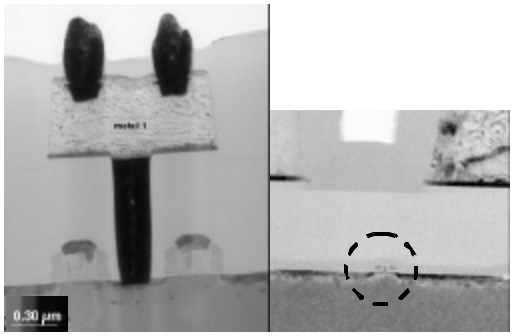


Figure 12: 0.18um technology node transistors (left), with 10nm transistor (circled on right) on the same scale

The transistor characteristics of a 10nm L_g device are shown in Figure 13. It can be seen that at this gate length, the transistor still behaves as a MOS device, although there is now increased conductance in the saturation region, and the leakage current (at V_g=0V) continues to increase. This is in part due to a relatively thicker gate oxide used in the present study than required at this technology node. A thinner Tox would give much better Short Channel Effects (SCE) control.

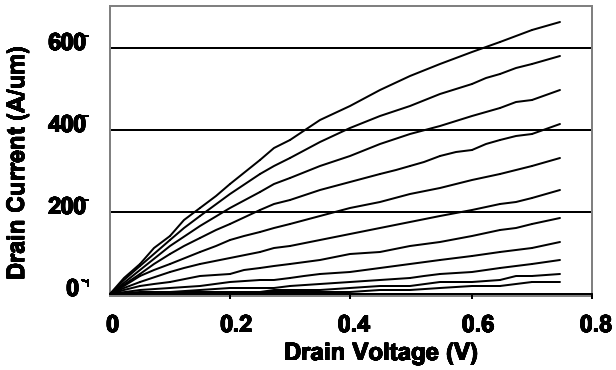


Figure 13: Id-Vd curves of 10nm transistor. V_g to 0.75V, steps of 0.1V

It should be noted that the supply voltage has been scaled in going from 30nm to 10nm poly lengths. Nevertheless, the leakage for the smallest gate length transistors is an issue of some importance. This is discussed later.

Taking the drive current values for the gate delay and energy delay, and plotting these against published data for longer gate lengths (Figures 14 and 15), it can be seen that the devices continue to scale on the same historical rate, even down to the lowest gate lengths. At 15nm, the gate delay is 0.39psec, and for 10nm gate length, the gate delay has dropped to 0.11psec (Figure 14). Similarly, the energy-delay product also drops exponentially, as can be seen in Figure 15, decreasing almost two orders of magnitude between the 30nm transistor and the 10nm transistor. Thus, even though the drive currents on these research transistors are not high (due to voltage scaling), they maintain the historic trend in gate delay and energy-delay.

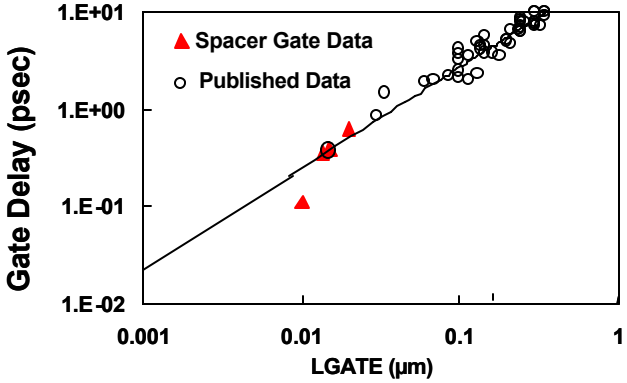


Figure 14: Gate delay for published & Intel Spacer Gate transistors

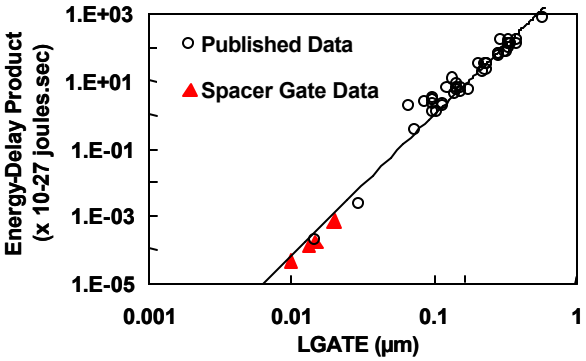


Figure 15: Energy delay for published & Intel Spacer Gate transistors

TRANSISTOR LEAKAGES

Junction Leakage

Returning to the issue of leakage current, there are three dominant sources of leakage: junction leakage, gate leakage, and off-state leakage. These three sources of leakage increase as transistors are scaled down towards 10nm.

Commencing with junction leakage, it has been suggested that this source of leakage alone will limit scaling [6]. This leakage arises from the high doping concentration in the channel region required to attain threshold voltages, and to limit short channel effects in aggressively scaled devices. The proximity of the valence and conduction bands in the depletion region of the junctions causes a parasitic tunneling current. Figure 16 shows the junction

edge leakage (I_{JE}) as a function of substrate doping at 25°C and 1V reverse bias. Although the leakages are high (above 1nA/ m at L_g=30nm), they are still a lot less than the other sources of leakage at 30nm, with less than 1.0nA/um for both n-MOS and p-MOS, a small percent of the transistor I_{off}.

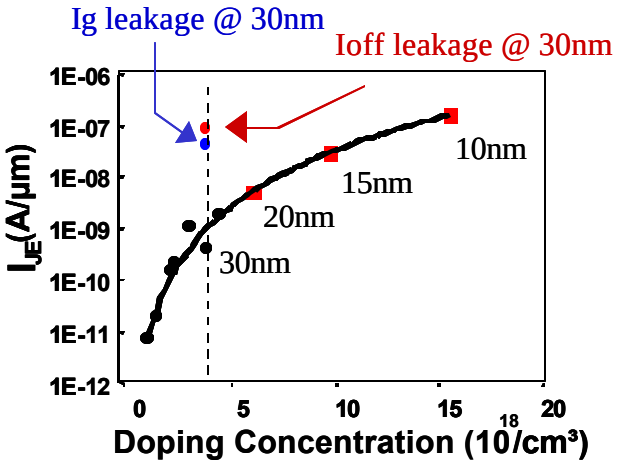


Figure 16: Junction leakage vs doping concentration. Circles - data, squares - extrapolated points. Other sources of leakage at L_g=30nm have been added to the graph

The arrows in Figure 16 indicate gate leakage and off-state leakage for L_g=30nm, both of which are more than an order of magnitude greater than the junction leakage. For the shorter channel devices, extrapolating to the 10nm gate lengths, and assuming a 1.6x doping concentration increase per technology generation, the junction leakage is still far below a value of 1 A/ m, the upper leakage limit.

Gate-Oxide Leakage

With respect to other sources of leakage, gate-oxide scaling has long been considered an eventual limiter for gate oxides below ~2nm gate dielectric thickness [7]. It was felt that with oxides reaching the thickness of several atoms, gate leakage would rival and would surpass the transistor off-current leakage. However, the examination of gate oxides down to 0.8nm [5] has not shown this to be the case in the present study. Figure 17 shows the gate current versus gate bias for the 0.8nm oxides [5]. The measurement results show that at 0.85V and 100°C, the gate leakage value is in the mid-10^{-8}A/ m^2, approaching the off-state leakage level of the 30nm L_g transistor (see Figure 16).

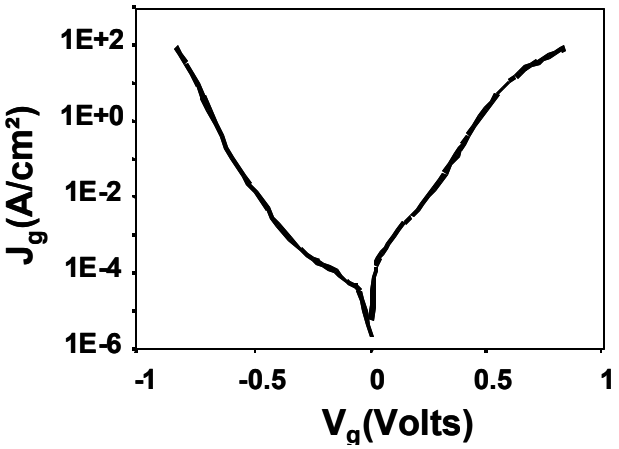


Figure 17: Gate current leakage for a 0.8nm oxide for the 30nm transistor [5]

Extrapolating further, below 0.8nm of gate-oxide thickness, leakage will become a limiter. With this in mind, research on high-K dielectrics for MOS transistor applications has become an area of active research. The reason for this is shown in Figure 18. It can be seen here that for the same equivalent oxide thickness (the thickness that SiO_2 would have for a given capacitance value), the high-k dielectric has more than four orders of magnitude less gate leakage than SiO_2.

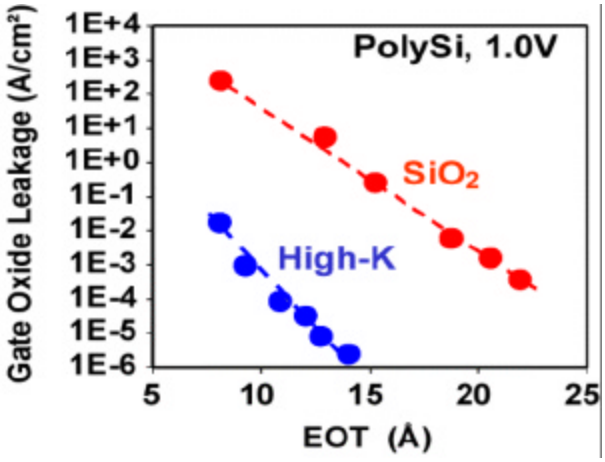


Figure 18: Comparison of gate leakage between SiO_2 and high-K dielectrics

Thus, for future scaling, a change in the transistor architecture to include high-K dielectrics will be necessary if gate capacitance scaling is to continue down to 10nm gate lengths.

Off-Current Leakage

Transistor leakage is perhaps the greatest problem facing continued scaling. As the transistor scales, the internal fields become greater, which necessitates scaling of the power supply voltage. This is also driven by the need to

decrease the power (P) generated by the chip, which is governed by the equation

$$P=C.V^2.f + I_{off}.V \qquad (1)$$

where f is the chip frequency, C is the gate-oxide capacitance, and I_{off} is the total transistor leakage current for the chip. The first part of Equation 1 refers to the active power, and the second refers to the off-state power. From this equation, it can be seen that reducing V has a significant effect on power. Scaling the power supply also necessitates the scaling of V_t, if Idsat is to be maintained. However, decreasing V_t results in increasing I_{off}, which in turn increases the off-state power.

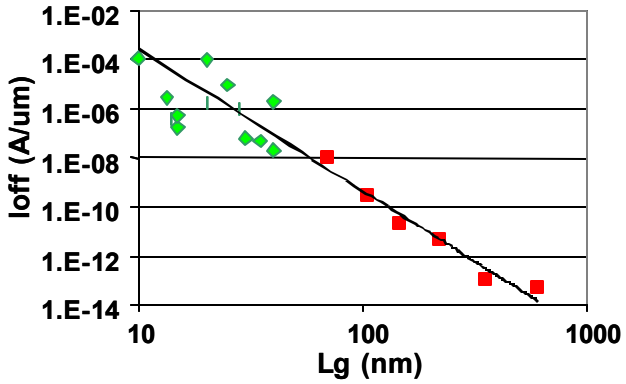


Figure 19: Transistor off-state leakage vs gate length
Red squares indicate pre-production transistors
Green diamonds indicate research devices

Figure 19 shows transistor off-state (source-drain) leakage versus transistor gate length. The red symbols are taken from the literature for transistors near production at the time of publication, while the green symbols are for 'research' devices (devices several generations from production at the time of their publication). Drawing a trend line through the data shows that the research transistors fall roughly on the same trend line that the advanced production devices fall on, irrespective of the gate length, obeying the power law relationship:

$$I_{off}=A.L_g^{-5.6} \qquad (2)$$

The fact that the research devices below 50nm follow the same relationship established for *pre-production* transistors at longer gate lengths suggests that this relationship is intrinsic to the scaling of bulk transistors using current methodologies (meaning the scaling of gate length, drive current, and voltage at the same time). Historically, it is the leakage current that has been relaxed to enable us to achieve the drive current scaling. If this same scaling methodology continues, controlling off-currents while at the same time maintaining aggressive drive currents will be difficult in bulk CMOS.

The effect of increasing I_{off} on total power [8] is illustrated in Fig. 20. In this figure, the off-state and active power components to the total leakage are plotted by taking 30 meters of transistor width for each technology generation (note that the chip leakage per generation increases since the total transistor width per generation also increases).

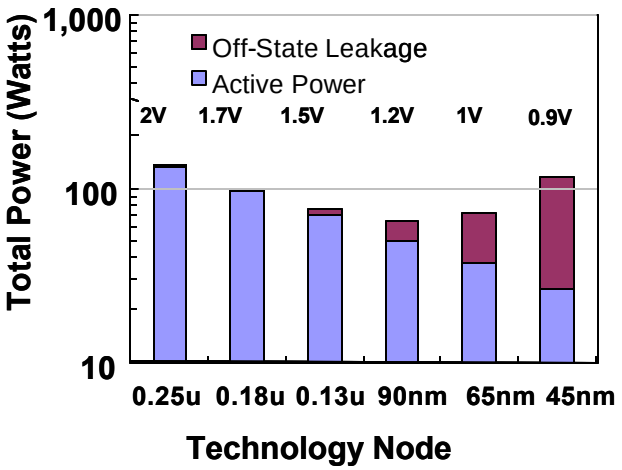


Figure 20: Total power as a function of technology node, for a fixed (30m) total transistor width, showing the increasing importance of off-state current as technology scales

It can be seen that the off-state leakage component to total power exceeds active power as the technology decreases beyond the 65nm node. Thus, finding a solution to the off-current issue is one of the most important challenges facing transistor design.

TRANSISTOR ARCHITECTURE FOR LEAKAGE

Depleted Substrate Transistor–Single Gate

One of the ways to overcome the issue of static power is to increase the threshold voltage. However, increasing the threshold voltage while scaling the power-supply voltage decreases the drive current of the device. A feasible way of addressing the power issue is to improve the sub-threshold gradient of the transistor. As the transistor scales, and the channel doping increases to support the thinner oxide, the sub-threshold gradient degrades. The sub-threshold gradient is linked to the depletion capacitance by the equation

$$S=(kT/q).\ln 10.(1+C_D/C_{ox}) \qquad [4]$$

where T is the temperature, q is the electronic charge, S is the sub-threshold gradient, CD is the capacitance of the depletion region, and C_{ox} is the gate-oxide capacitance [9]. From equation 4, it can be seen that decreasing the depletion capacitance C_d will improve the sub-threshold

gradient towards the minimum theoretical value of 60mV/decade. Decreasing C_D can be achieved by the use of Silicon-On-Insulator with a fully depleted substrate, since the depletion layer now extends through the buried oxide into the substrate. The value of C_D then becomes negligible compared to C_{ox} in Equation 4. We call this broad category of devices, which include several elements necessary for future scaling, a Depleted Substrate Transistor (DST), and this can be seen in Figure 21 [10].

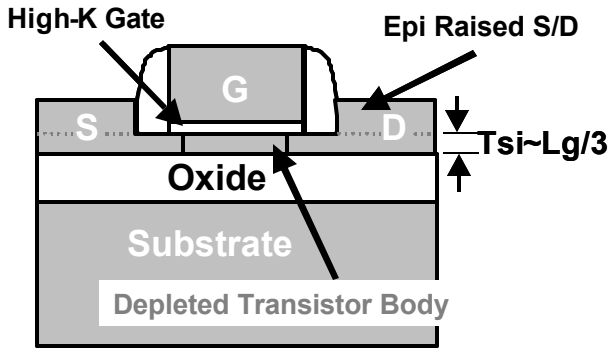


Figure 21: Illustration of Depleted Substrate Transistor (DST)

We define the DST as consisting of three elements:

The body of the device is fully depleted, be this double-gate (DG) [9] transistors, gate-all-around transistors (GAA) [11], or even transistors whose bodies are no longer silicon (III-V's etc.).

The gate dielectric is a high-k material mentioned previously.

The junctions are raised epitaxial source/drain.

Figure 22 shows a TEM cross section of such a device. The epitaxial raised source/drain are necessary to decrease the series resistance of the transistor, due to the thin silicon body.

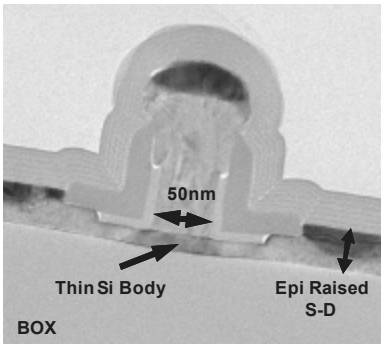


Figure 22: TEM of the Depleted Substrate Transistor (DST)

Depleted Substrate CMOS transistors were fabricated on a thin silicon body with a thickness of <25nm on top of a ~200nm buried oxide. The physical gate-oxide thickness was equal to 1.5nm, the same as the bulk devices. Figure 23 shows the I_d-V_g characteristics of two 60nm L_g n-MOS transistors, a bulk transistor and a DST.

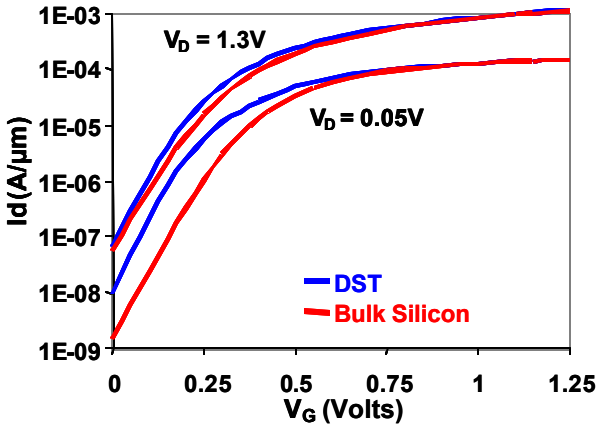


Figure 23: Log Id-V_g characteristics of a bulk and DST transistor

Two features can be seen from this figure: the sub-threshold gradient, which has improved from 95 mV/decade to 75mV/decade, and the DIBL, which has decreased from 100mV/V to 45 mV/V. The improved sub-threshold gradient thus allows the DST to decrease threshold voltage by 40-50mV, while the improvement in DIBL allows a further 50mV decrease in V_t. As power supply voltages decrease below 1V, the DST devices will allow substantial gains in gate overdrive (V_g-V_t), as well a reduction in off-state power by 2 orders of magnitude.

Depleted Substrate Transistor–Double Gate

As transistors continue to scale, control of short channel effects become more and more important. It will be increasingly difficult to control the electrostatic communication between source and drain that results in transistor leakage by using bulk or even single-gate DSTs. Solutions are being researched that enclose the channel area by the gate stack. The most common form of this transistor architecture is called the *Double-Gate* transistor.

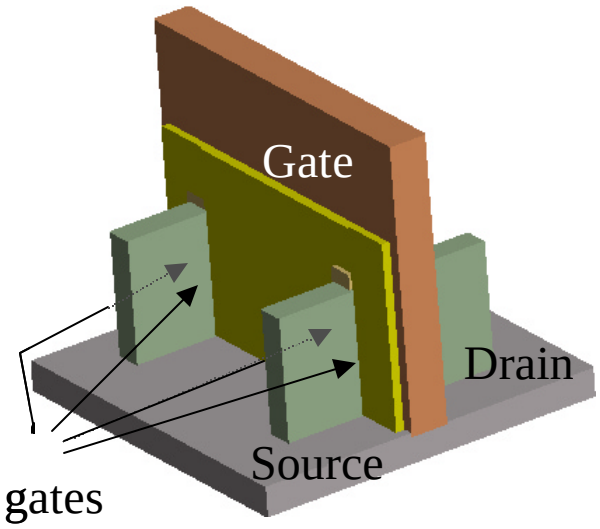


Figure 24: Illustration of a two-FinFET Double-Gate transistor. The current flows along each sidewall of the fins

Figure 24 shows an illustration of a two-fin transistor, one form of double-gate device. In this case, the current runs along both sidewalls of the fins.

The gate controls the front and back of such a double-gate transistor, thus offering better short channel control than a single gate. However, double-gate devices are much more difficult to fabricate due to their three-dimensional nature. Figure 25 shows a double-gate FinFET device in the direction of current flow and Figure 26 shows the transistor perpendicular to current flow.

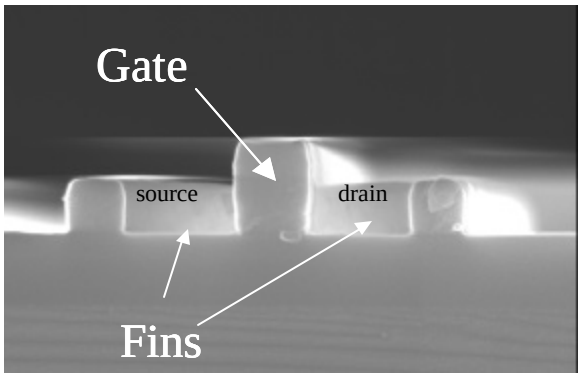


Figure 25: SEM of Double-Gate Multi-Fin Structure

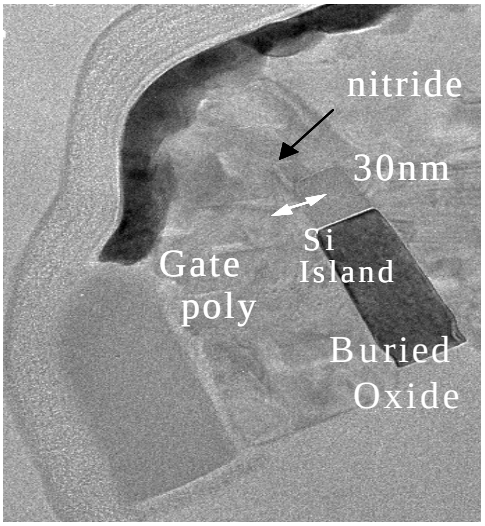


Figure 26: TEM cross-section of a 30nm double-gate device

In terms of short channel control, simulations have shown that double-gate devices can buy up to a two-generational gain in DIBL over single-gate DSTs [12]. However, it should be noted that one of the issues concerning both types of device is the thickness of the silicon that forms the channel region. In the case of single-gate DSTs, the thickness of the silicon channel body has been found to be approx $L_g/3$. In the case of double-gate DSTs, the thickness of the Fin is twice the body thickness ($2L_g/3$), as each gate controls a thickness of $L_g/3$.

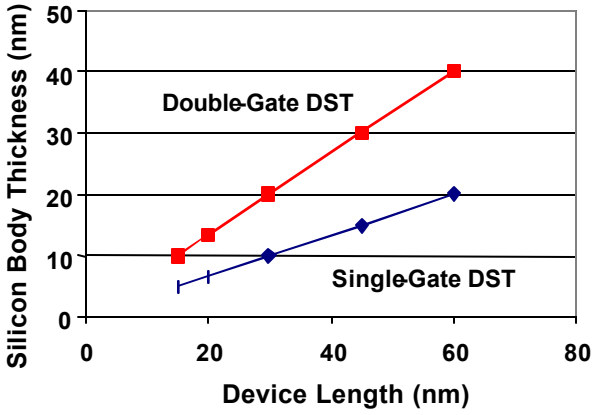


Figure 27: Silicon body thickness required for full depletion of a single-gate DST and a double-gate DST

As the gate length scales, the thickness of the silicon body (Tsi) also scales. Figure 27 shows the thicknesses needed to provide full depletion for both single- and double-gate DSTs. It can be seen that for single gates, the thickness quickly approaches to less than 10nm. This constraint of requiring Tsi < 10nm is relaxed in FinFETs,

since the Tsi is perpendicular to the wafer plane (Figure 24), and the thickness values are twice that of single-gate DSTs (Tsi is the fin width in the case of double-gates). However, this dimension is achieved using lithography, and this means that the most critical lithography step is no longer polysilicon patterning, but Fin patterning. In other words, for FinFET devices, the fin width needs to be smaller than the gate length. For example, for 20nm L_gs, the Fin patterning will require lithography that can reproducibly print 13nm Fin widths.

Drive Current

One of the most serious issues with gate length scaling is our ability to maintain high drain current as the power-supply voltage scales without being able to fully scale V_t, which remains high to control transistor leakage currents. The power-supply scaling shown in Figure 1 suggests that keeping Idsat constant will be a significant challenge. In order to illustrate this point, Figure 28 shows the data from a 20nm gate length device at Vdd=0.85V and Vdd=0.7V. It can be seen that a power- supply voltage drop of 0.15V results in a drop of 30% in the drive current capabilities of the transistor, from 533 A/ m to 375 A/ m. Some of the issues facing drive current scaling are discussed below.

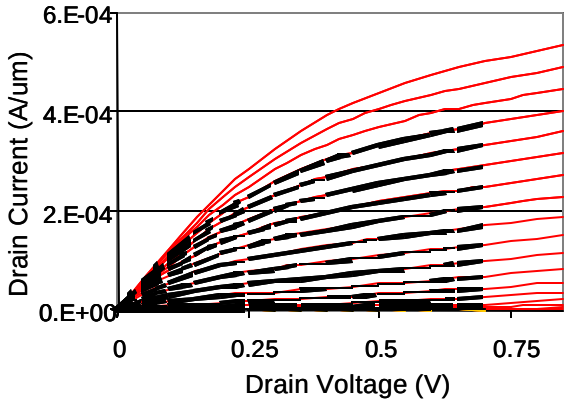


Figure 28: Id-Vd characteristics for the 20nm transistor at two different supply voltages, 0.85V and 0.7V

Series Resistance

With DST-like devices comes the need to keep the body thickness in the range that allows for complete depletion. In the representation of the DST transistor in Figure 21, the thickness of the silicon body (T_{Si}) needs to be kept to around $L_g/3$ to maintain complete depletion (see also Figure 27). As the transistors scale to L_g=20nm, the body thickness will need to be of the order of 6-7nm. Apart from the fabrication issues for such thin bodies discussed above, the increase in series resistance arising from ultra-thin junctions will limit transistor drive currents [13].

The solution to the drive current issue (from external parasitic resistance) is to use raised source/drain, which increases the effective thickness of the junctions and hence the junction conductance [14]. Figure 29 shows the advantage of raised source/drains over conventional junctions on DST transistors. The transistors with and without raised source/drain were fabricated with a gate length of 60nm. At matched I_{off} of 60nA/îm, DST devices with raised source/drain (blue lines) exhibit superior drive currents, up to 50% more than the non-raised source/drain DST structures (red lines).

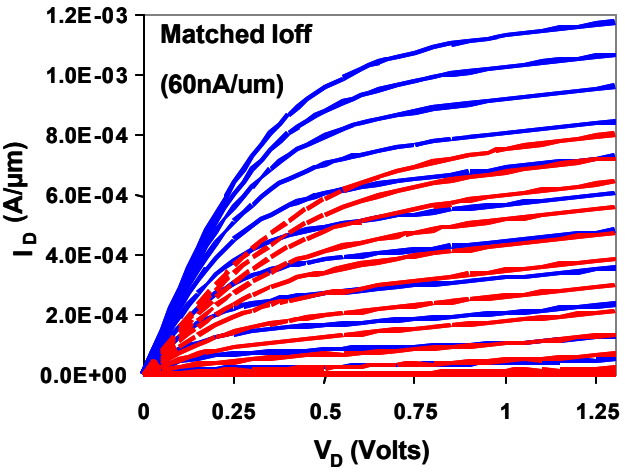


Figure 29: I_D-V_D characteristics for 60nm DST transistors, with no raised S/D (red lines), and with raised S/D (blue lines)

Figure 30 further illustrates the performance gains that can be obtained in combining DST with raised source/drains. Figure 30 shows the PMOS I_{on}-I_{off} comparison of the depleted-substrate transistor with and without raised source/drain, and the standard 0.13um-generation bulk Si transistors at V_d = 1.3V. For a given I_{off} (e.g., 1.0 nA/um), the depleted-substrate transistor with raised source/drain shows the highest I_{on} value, about 30% higher than the standard bulk Si transistor. Conversely, at a fixed drive current (e.g., at 0.6mA/ m), the off-current is decreased by about two orders of magnitude for DST.

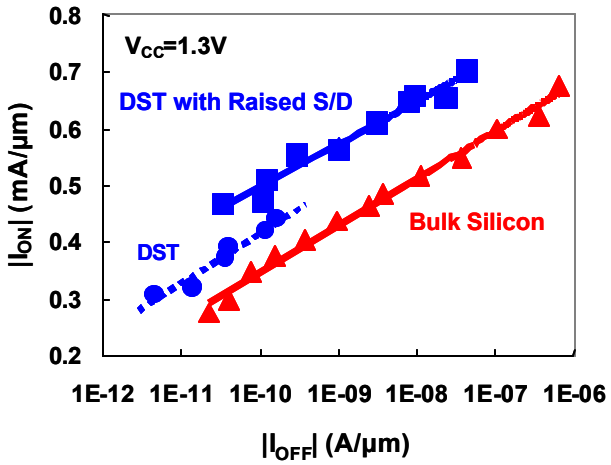


Figure 30: Comparson of p-MOS bulk silicon (triangles), DST (circles) and DST with raised source/drains (squares)

Another way of looking at the data is from a power-supply perspective. DST pMOS with raised source/drain achieves the same I_{on}-I_{off} performance at 1.1V as the bulk-device at 1.3V, thus enabling a reduction in power by 30% (power voltage^2).

GATE STACK

As discussed in a previous section, future transistor design will need to incorporate high-K dielectrics for continued transistor scaling. One of the considerations with high-k dielectrics is the dielectric integrity at high frequencies. With clock speeds already in the gigaherz, the gate material must maintain its dielectric integrity to frequencies well above this. If the responding material cannot follow the switching at high frequencies, the high dielectric constant measured normally at low frequencies will not be obtained, and therefore the gate capacitance and subsequent drive current will be reduced. Figure 31 shows measurements of dielectric constant as a function of frequency for three different materials, SiO_2, HfO_2, and ZrO_2, up to 20GHz. It can be seen that HfO_2 and ZrO_2 show the same invariance to frequency that the SiO2 dielectric shows.

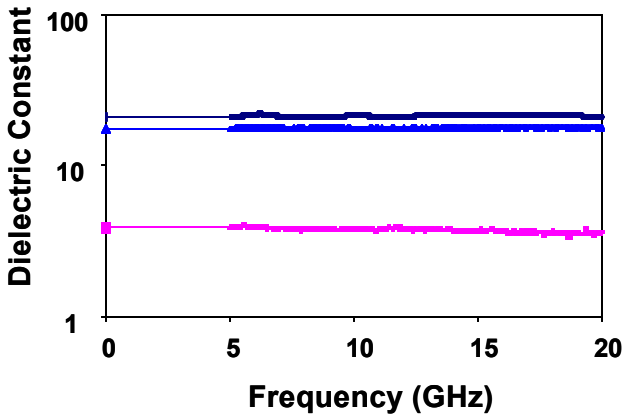


Figure 31: Dielectric constant versus frequency for SiO_2 (bottom) and the High-Ks, ZrO_2 and HfO_2 (bottom)

The dielectric itself is not, however, the only issue in maintaining high gate capacitance. There are other capacitance elements that tend to reduce the net gate capacitance. The gate stack enters into the transistor saturation current equation through the term C_{oxe}:

$$Id = .C_{oxe}.(Z/L).(V_g-V_t)^2 \qquad [3]$$

Where C_{oxe} is the equivalent capacitance of the gate stack (C_{oxe}= $_{SiO2}/T_{oxe}$), and is made up of

$$T_{oxe}=T_{oxp} + T_{qm} + T_{pd} \qquad [4]$$

where T_{oxp} is the equivalent thickness of the dielectric itself, if it were SiO2; T_{qm} is the quantum mechanical term coming from quantization of the inversion layer, which tends to cause the electrons to reside in the silicon a short distance away from the Si/SiO_2 interface; and T_{pd} is the depletion region in the poly electrode resulting form incomplete degeneration of the gate electrode. The values typically taken for these are 0.5nm for poly depletion and 0.5nm for quantum mechanical effects [15].

The quantum mechanical contribution always persists, and even if dielectric thickness is reduced to zero, the electrical T_{ox} (inversion) would approach 1nm. A further increase in capacitance can be achieved by eliminating the poly depletion portion with use of a conductive metal gate electrode. This would reduce the equivalent oxide thickness by 0.5nm. This approach is being actively researched (e.g., [16]).

CONCLUSIONS

Transistor scaling issues have been examined to determine the implications on device performance. We have fabricated planar Si transistors down to 10nm physical gate length using a special spacer gate technique. Transistors at these aggressively scaled dimensions down

to 15nm are shown to exhibit good device characteristics. Although transistors with 10nm physical gate length show normal switching characteristics, they exhibit very high off-state leakage. To alleviate the high parasitic leakage problem, we have demonstrated a transistor structure with a fully depleted substrate (DST) providing near-ideal sub-threshold gradient and highly reduced DIBL. In addition to DST device architecture, new electronic materials and modules will be needed in the future to maintain high performance and low parasitic leakages.

REFERENCES

[1] ITRS Roadmap, "Process Integration, Devices and Structures and Emerging Research Devices section, 2001 Edition.

[2] H. Liu et al., "A Patterning Process with sub-10nm 3-CD Control for 0.1 m CMOS Technologies" *SPIE* vol. 3331, pp. 375-381, 1997.

[3] Y-K Choi et. al., "A Spacer Patterning Technology for Nanoscale CMOS," *IEEE Trans. Electron Device Letters*, vol. 49, pp. 436-441, 2002.

[4] D. W. Barlage et. al., "Inversion MOS capacitance extraction for high-leakage dielectrics using a transmission line equivalent circuit," *IEEE Electron Device Letters*, vol. 21, pp. 454-456, 2000.

[5] R. Chau et. al., "30 nm physical gate length CMOS transistors with 1.0 ps n-MOS and 1.7 ps p-MOS gate delays" *IEDM* , 2000, pp. 45-48.

[6] T. Ghani et. al., "Scaling challenges and device design requirements for high-performance sub-50 nm gate length planar CMOS transistors," *VLSI*, 2000, pp. 174-175.

[7] J.H. Stathis et. al., "Reliability projection for ultra-thin oxides at low voltage," *IEDM* 1998, pp. 167-170.

[8] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," *1999 ISLPED*, pp. 163-168, August 1999.

[9] J-P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI*, Kluwer Academic Publishers, 1991.

[10] R. Chau et. al., "A 50nm depleted-substrate CMOS transistor (DST)," *IEDM* 2001, pp. 621-624.

[11] J-P. Colinge et. al., "Silicon-On-Insulator Gate-All-Around Device," *IEDM*, 1990, pp. 595-599.

[12] H.-S.P. Wong, "Device design considerations for double-gate, ground-plane, and single-gated ultra-thin SOI MOSFET's at the 25 nm channel length generation," *IEDM* 1998, pp. 407-410.

[13]Kim, S.D., et. al., "Advanced model and analysis of series resistance for CMOS scaling into nanometer regime," *IEEE Trans. Electron Devices*, Vol. 49, pp. 457-472, 2000.

[14] S. Yamakawa et. al., "Drivability improvement on deep-submicron MOSFETs by elevation of source/drain regions," *IEEE Electron Device Letters*, Vol. 20., pp. 366-368, 1999.

[15] S.V. Walstra et. al., "Thin oxide thickness extrapolation from capacitance-voltage measurements," *IEEE Transactions on Electron Devices*, Vol. 44., pp. 1136-1142, 1997.

[16] I. Polishchuk et. al., "Dual work function metal gate CMOS transistors by Ni-Ti interdiffusion," *IEEE Electron Device Letters*, Vol. 23, pp. 200-212, 2002.

AUTHORS' BIOGRAPHIES

Brian Doyle joined Components Research in Intel Corporation in 1994, after working for Digital Equipment Corporation and Bull S.A. He worked on new technology modules in Santa Clara before moving to Oregon in 1999. His primary focus is on new transistor architectures. He received his B.Sc. from Trinity College, Dublin, and his M.S. and Ph.D. degrees from the University of London. His e-mail is brian.s.doyle@intel.com.

Reza Arghavani graduated in 1991 from UCLA with a Ph.D. degree in Solid State Physics. He has worked in the fields of fault isolation, gate dielectrics and device engineering at Intel. His e-mail is reza.arghavani@intel.com.

Doug Barlage is currently working on advanced CMOS for Intel Corporation. His primary focus has been on dielectric characterization. He has established techniques for assessing the capacitance in leaky dielectrics as well as determining the dielectric roll-off with respect to frequency for thin gate dielectrics. He received his Ph.D. degree from the University of Illinois in GaAs based devices for high speed and millimeter-wave circuits. His e-mail is douglas.barlage@intel.com.

Suman Datta has been at Intel for over two years working on logic transistor design and development. His main interests are in advanced gate stack engineering and new transistor architectures. Suman received his B.S. degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, in 1995, and his Ph.D. degree in Electrical Engineering from the University of Cincinnati, Ohio, in 1999. His e-mail is suman.datta@intel.com.

Mark Doczy joined the Components Research group at Intel in 1996 after receiving his Ph.D. degree in Plasma Physics from the University of Wisconsin. Upon joining Intel, Mark worked on plasma induced damage to transistors including gate oxide charging and line edge roughness. Mark is currently focused on novel device development. His email is mark.doczy@intel.com.

Jack Kavalieros has been with Intel for seven years. He received his Ph.D. degree from the University of Florida in 1995. He is responsible for novel device process integration as well as novel gate oxide development. His e-mail is jack.t.kavalieros@intel.com.

Anand Murthy received his Ph.D. degree from the University of Southern California and joined Intel in 1995. His current focus is on novel materials deposition for transistor research. His e-mail is anand.murthy@intel.com

Robert Chau is an Intel Fellow and Director of Transistor Research in the Logic Technology Development group, responsible for directing research and development in advanced transistors and gate dielectrics for microprocessor applications. Robert currently manages the Novel Device Laboratory and leads a research team focusing on new transistor architectures, process modules and technologies, and characterization techniques for the 65nm and 45nm logic technology nodes and beyond. With a B.S., M.S., and Ph.D. in Electrical Engineering from the Ohio State University, Robert holds 26 US patents and has received four Intel Achievement Awards and 13 Intel Logic Technology Development Division Recognition Awards for his outstanding technical achievements in research and development. His e-mail is Robert.s.chau@intel.com

Copyright © Intel Corporation 2002. This publication was downloaded from http://developer.intel.com/ .

Legal notices at
http://www.intel.com/sites/corporate/tradmarx.htm.

The Intel Lithography Roadmap

Peter J. Silverman, Technology and Manufacturing Group, Intel Corporation

Index words: Lithography, Moore's Law, 193nm, 157nm, EUV, Affordability

ABSTRACT

Lithography is the primary enabling technology for semiconductor manufacturing. Having led the industry transition to Deep Ultra-Violet (DUV) lithography, Intel is currently leading the transition to 193nm, 157nm, and Extreme Ultra-Violet (EUV) lithography. Lithography technologies, such as 193nm, 157nm, and EUV lithography, which have benefited from Intel investment, have gained industry acceptance, while competing technologies, such as x-ray lithography, are no longer being pursued.

The Intel Lithography Roadmap is the Intel plan for the next several generations of lithography technology. In this paper, we discuss this roadmap and review the strategic and tactical forces that have produced the current version of this roadmap. The status of future lithography technologies is also reviewed, with an emphasis on 193nm, 157nm, and EUV lithography. Finally, the key question of affordability is addressed.

INTRODUCTION

Lithography is the single most important driver of Moore's law. By providing the capability to continuously reduce the size of features patterned on semiconductor wafers, each new generation of lithography equipment has enabled faster microprocessors and smaller, less costly integrated circuits. Without the continuous improvements in lithography process and equipment technology that have occurred over the past 30 years, personal computers, cell phones, and the Internet would not be available today.

Due to the importance of lithography, Intel devotes large amounts of time and money to developing a strategic and tactical roadmap for the future direction of Intel lithography technology. Because the semiconductor industry has aligned with the Intel Lithography Roadmap, Intel's decisions have a strong influence on the investment decisions made by the suppliers who provide lithography equipment to the semiconductor industry. For example, Intel leadership was the catalyst for industry investment in 157nm lithography. Similarly, Intel has been the force behind the semiconductor industry acceptance of EUV lithography as the successor to traditional optical lithography. The strong influence of the Intel Lithography Roadmap makes it worthwhile to review both the roadmap and the forces that have created it.

The Intel Lithography Roadmap is driven by technical forces such as lithographic resolution and process control; tactical forces such as the development schedule for new lithography equipment; and commercial forces such as the affordability of lithography equipment. A review of these forces explains how the current Intel Lithography Roadmap has been developed. Intel has made the decision to invest in certain lithography technologies, such as 193nm, 157nm, and EUV lithography and not to invest in other technologies, such as x-ray lithography and electron projection lithography. A review of the status and timing of future lithography technologies provides insight into the decisions Intel has made in the past and will make in the future.

It is well known that the cost of lithography equipment has increased at a nearly exponential rate over the past 30 years. The $100,000 contact printers of the early 1970s have given way to the over $12M 193nm step-and-scan exposure tools of the first decade of the 21st century. How will the semiconductor industry be able to afford such costly equipment? Will the investment in future lithography technologies be wasted because of other limitations on transistor size reduction (transistor scaling)? Intel has a high degree of confidence in the ability of the industry to continue transistor scaling for many years into the future. Furthermore, Intel has a well-developed strategy to manage lithography affordability. This strategy will enable continued transistor scaling.

This paper reviews the strategic decisions and thinking that have resulted in today's Intel Lithography Roadmap. The status of advanced technologies such as 193nm, 157nm, and EUV lithography are reviewed in order to provide the background for Intel's roadmap decisions. Information is presented to support the continuing need

for advanced lithography technologies to enable transistor scaling. Finally, the affordability of future lithography technologies is addressed.

INTEL LITHOGRAPHY ROADMAP

The Intel Lithography Roadmap is the plan for the lithography technology that will be used to pattern the smallest features on each new generation of integrated circuits. Contemporary semiconductor devices have ~25 patterned layers. The smallest features are on the four to six "critical" layers, which define the size of the transistors. The remaining layers are used to interconnect the transistors to form an integrated circuit. Interconnect layers have larger feature sizes. As discussed in the section on affordability, the interconnect layers are normally patterned by "reusing" lithography equipment from earlier process generations.

Intel always uses the most advanced lithography technology that is ready for manufacturing to pattern critical layers. As shown in Figure 1, Intel is using DUV (248nm) lithography for the critical layers of the 130nm generation. The Intel plan is to transition to 193nm lithography for the 90nm generation; 157nm will be used on the critical layers of the 65nm generation if 157nm lithography is ready on time, and 157nm lithography will be used on the critical layers of the 45nm generation if EUV is not ready on time.

The dates shown in Figure 1 are for the start of high-volume manufacturing. However, lithography tools for process development are required at least two years sooner. Furthermore, equipment suppliers require approximately five years to design and build each new generation of lithography equipment. Therefore, the ten-year look-ahead provided by the roadmap is needed to allow both Intel and the equipment suppliers to plan for the future.

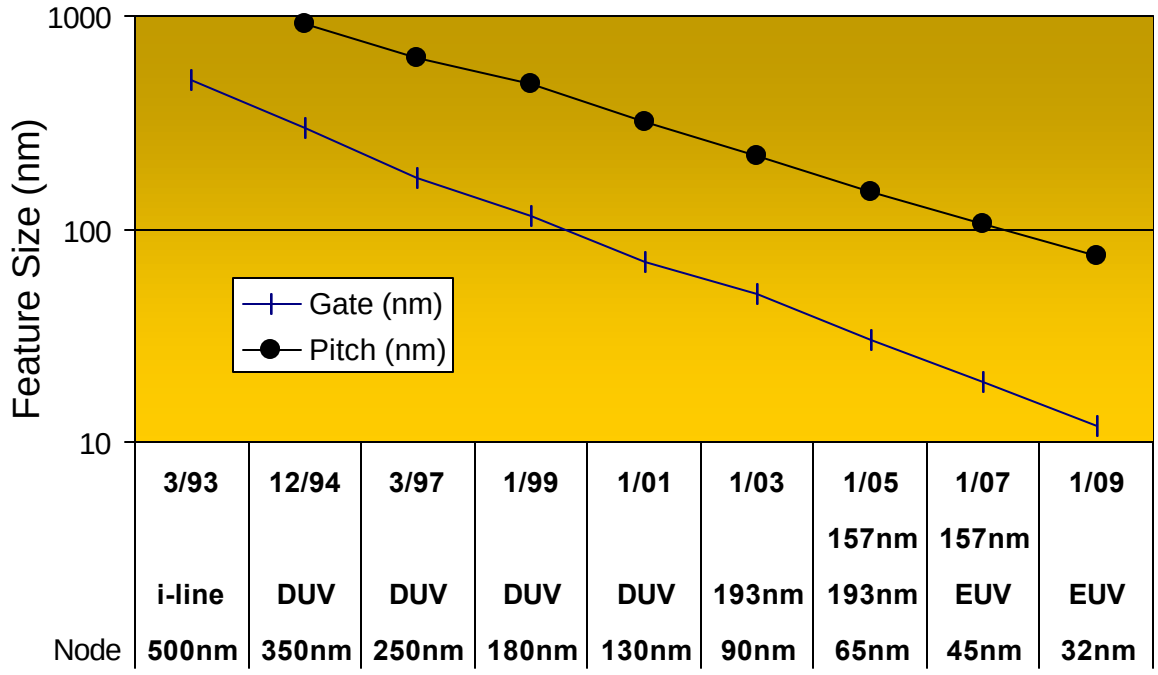


Figure 1: Intel Lithography Roadmap

The Intel Lithography Roadmap shows a continuous progression to shorter lithography wavelengths (smaller). Starting with i-line (365nm) lithography, the roadmap progresses to DUV (248nm), 193nm, 157nm and EUV (13nm) lithography. The drive to shorter wavelengths is because optical resolution is directly proportional to wavelength. Using a shorter wavelength enables manufacturing integrated circuits with smaller transistors.

Strategic and Technical Drivers

For nearly 30 years the growth of the semiconductor industry has been tied to Moore's Law; the essence of which is the ability to give customers faster, more complex products by manufacturing faster, more complex

integrated circuits, at a constant or decreasing price. The Intel Lithography Roadmap is driven by a commitment to maintain the industry momentum provided by Moore's Law.

In lithographic terms, Moore's Law translates into three technical requirements:

1. *Reduce pitch by 30% every two years*. A 30% reduction in pitch produces a 50% reduction in chip area. This allows more complex products to be produced without an increase in chip size.
2. *Reduce gate width by >30% every two years*. Since transistor speed is inversely proportional to gate width, smaller gates mean faster chips.
3. *Maintain a constant cost for lithography*. Since lithography is the largest single component of chip fabrication cost, lithography costs must stay constant to allow chip costs to stay constant.

Figure 1 shows the 30%/generation pitch and gate size reduction, which Intel has maintained on a two-year cycle for the past ten years.

Intel's roadmap strategy is designed to ensure that these requirements are met for each new generation of Intel technology. Therefore, lithography decisions are based on staying on the two-year cycle of Moore's Law and on meeting device density and speed requirements with affordable lithography technology.

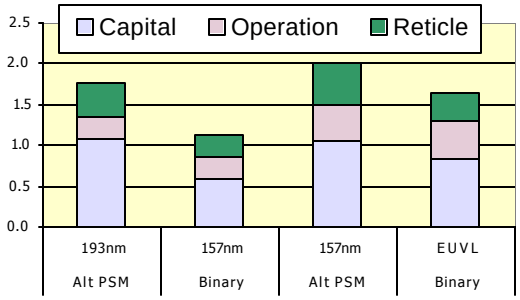


Figure 2: Binary vs. PSM cost/layer

Intel Roadmap Strategy

Semiconductor manufacturers follow two different roadmap strategies. Some companies work very hard to extend their existing, in-use lithography technology for as many generations as possible. Other companies transition as rapidly as possible to each new generation of lithography technology. Intel follows both strategies simultaneously. For the critical, transistor device layers, the Intel strategy is to transition as rapidly as possible to each new generation of lithography technology. For the less critical, interconnect layers, the Intel strategy is to reuse existing lithography equipment.

Intel transitions rapidly to new lithography technologies because we have found that this is the lowest total cost approach. Even though new generations of lithography equipment are more costly, the costs are more than offset by the savings in other areas; e.g., mask costs.

Figure 2 compares two potential candidates for the critical layers of the 65nm technology node (157nm with Binary masks and 193m with Alternating Phase Shift Masks) and two candidates for the 45nm node (EUV Lithography with Binary masks and 157nm with Alternating Phase Shift Masks). In both cases, the next-generation technology has significantly lower cost/layer due to less expensive masks and lower capital costs. (The lower capital costs are due to the higher run rate that is achievable with binary masks.) Therefore, Intel's plan is to use 157nm lithography on the 65nm node and to use EUV Lithography on the 45nm node. Of course, these plans are dependent on the availability of 157nm and EUV exposure tools in the required time frames.

Lithography Roadmap Acceleration

As shown in Table 1, i-line/g-line lithography was used for six technology generations over a period of fifteen years. DUV lithography will be used for three process generations over a period of six years. The Intel Lithography Roadmap (Figure 1) shows 193nm, 157nm, and EUV all being introduced in the following four years. What has happened to force the roadmap to accelerate so rapidly?

Table 1: Wavelength "Generations"

Year	Node	Lithography
1981	2000nm	i/g-line Steppers
1984	1500nm	i/g-line Steppers
1987	1000nm	i/g-line Steppers
1990	800nm	i/g-line Steppers
1993	500nm	i/g-line Steppers
1995	350nm	i-line → DUV
1997	250nm	DUV
1999	180nm	DUV
2001	130nm	DUV
2003	90nm	193nm
2005	65nm	193nm → 157nm
2007	45nm	157nm → EUV
2009	32nm and below	EUV

Two factors have contributed to the accelerated rate of change in lithography:

1. The transition to sub-wavelength patterning as shown in Figure 3.
2. The finite limit on the Numerical Aperture (NA) of optical systems, which sets a limit on the minimum possible resolution at a particular wavelength, as shown in Figure 4.

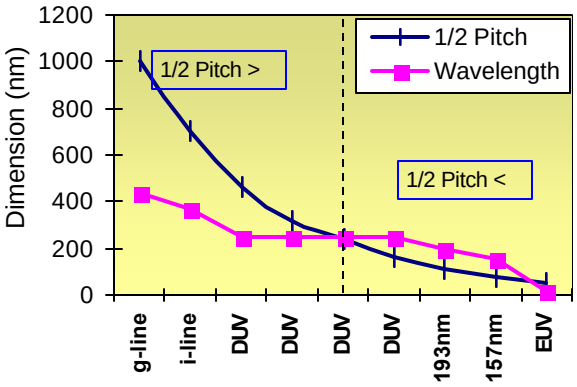


Figure 3: Sub-wavelength lithography

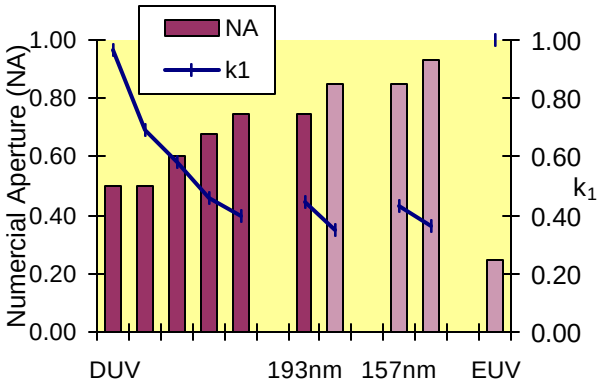


Figure 4: Limits of wavelength extension

Resolution is related to wavelength and NA by the well-known equation:

$$resolution \mu \frac{Wavelength}{Numerical\ Aperture}$$

The combined impact of these two factors has been to accelerate the rate of introduction of new lithography technologies; i.e., to accelerate the transition to ever smaller wavelengths. The need for smaller wavelengths to maintain Moore's Law is the primary reason that Intel has invested over $200M in the development of EUV lithography.

Transistor Scaling

Even if it is possible to use lithography to pattern features smaller than 50nm, there is legitimate concern as to whether other factors will constrain the ability of the semiconductor industry to manufacture 45nm generation and smaller transistors.

Intel has addressed this question by accelerating research on transistor design. The Intel announcement of TeraHertz transistors with gate dimensions below 20nm (Figure 5) clearly demonstrates that transistor physics and material properties will not prevent continuing on the path of Moore's Law. The key issue will be the availability of lithography equipment that can pattern sub-50nm features, in high-volume applications, at affordable costs. This again emphasizes the need to accelerate the lithography roadmap.

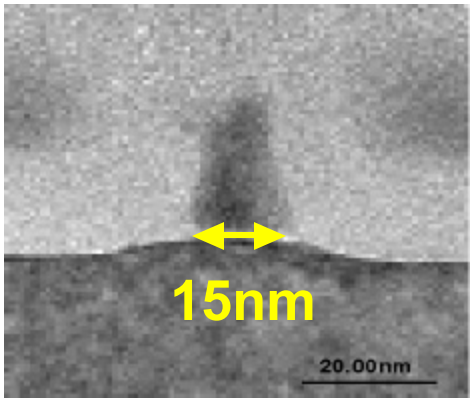


Figure 5: TeraHertz transistor with 15nm gate

TECHNOLOGY DEVELOPMENT STATUS

In the early 1990s, there was significant concern that the industry could not make the transition from i/g-line lithography to 248nm Deep Ultra-Violet (DUV) lithography. There were many challenges to overcome before DUV lithography could be successful in high-volume manufacturing. Exposure tool suppliers had to learn to fabricate precision optics from ultra-pure fused silica. Resist suppliers had to develop and commercialize chemically amplified resists. Mask makers had to learn to use new materials. However, all these challenges were overcome, and DUV (248nm) lithography has been the workhorse technology for semiconductor manufacturing since the 250nm (0.25 micron) generation. DUV exposure tools, which were introduced at 0.50NA, are now in their fourth generation, with fifth-generation, >0.80NA tools due in 2003.

The industry is poised to introduce 193nm lithography into high-volume manufacturing in the second half of 2002. Prototype 193nm exposure tools were delivered in 1996. Early production 193nm tools were delivered in 2001. Both the prototype and early production tools were delivered in small quantities, partly due to the lack of a mature 193nm resist technology. Mature, high-resolution 193nm resists are now available from several suppliers. Lithography equipment suppliers are ready to deliver production quantities of 0.75NA 193nm exposure tools in 2003 to support 90nm integrated circuit manufacturing on 300mm wafers. By early 2003, suppliers will be ready to deliver 0.85NA 193nm exposure tools to support development and early manufacturing of 65nm integrated circuits.

Patterning 65nm generation integrated circuits will require either 157nm lithography or 193nm lithography with Alternating Phase Shift Masks. Both Intel and the lithography equipment suppliers are confident that the cost of 157nm lithography will be less than the cost of 193nm lithography with Alternating Phase Shift Masks. There are many challenges to overcome before 157nm lithography can be used in high-volume manufacturing. The challenges include the development of large supplies of large diameter, high-purity CaF_2 crystals for optics; the development of pellicles with high transparency at 157nm to protect masks, and the development and commercialization of 157nm resists.

Although there are many challenges to 157nm lithography development, there has been excellent progress in the last few years. Suppliers have developed 157nm optical designs; materials for 157nm mask blanks are now available; and 157nm resists with good imaging capability have been demonstrated (Figure 6).

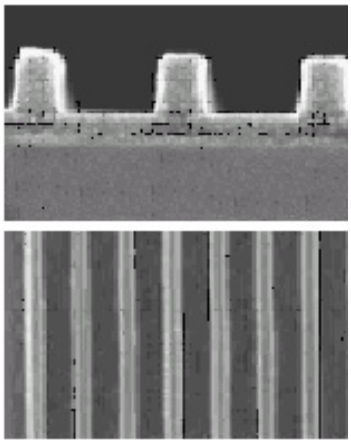


Figure 6: 157nm resist images (80nm lines)

The current forecast is that 157nm exposure tools will not be available until 2004. Therefore, 65nm integrated circuit technology will be developed using 193nm lithography. It is likely that 193nm lithography will also be used for early 65nm generation production. However, 157nm is expected to intersect the peak of the 65nm integrated circuit generation.

Extreme Ultra-Violet (EUV) lithography is being developed for 45nm generation integrated circuits. There has been excellent progress on EUV lithography in the past two years. The feasibility of manufacturing EUV optics has been demonstrated. EUV masks have been produced by several mask shops. EUV resists are available, since DUV resists are capable of EUV imaging. The EUV LLC (Limited Liability Company) has demonstrated that all the components of EUV technology can be integrated into a fully functional, 0.10NA, prototype EUV exposure tool (Figure 7), which can pattern 70nm features (Figure 8). The success of the prototype tool demonstrates that sub-50nm lithography will be possible with first-generation, production EUV exposure tools and that ~20nm lithography should be possible with second-generation EUV tools.

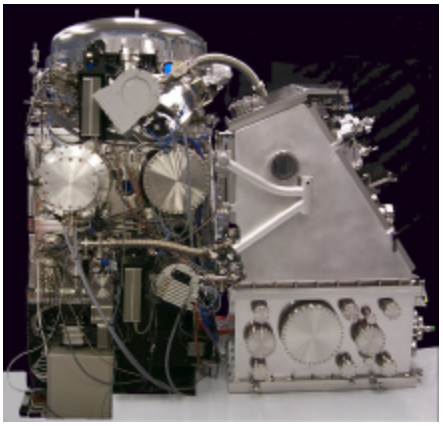


Figure 7: Prototype EUV exposure tool

However, there are still significant risks which could delay the introduction of EUV. For example, the lack of a high-power source of EUV radiation could reduce the run rate (output) of EUV exposure tools and make EUV too expensive for high-volume manufacturing. Thus, even with the excellent progress on EUV lithography, which has occurred over the past two years, the situation at the 45nm node is similar to the situation at the 65nm node.

Although there is a strong consensus that EUV lithography will be used at the 32nm generation and

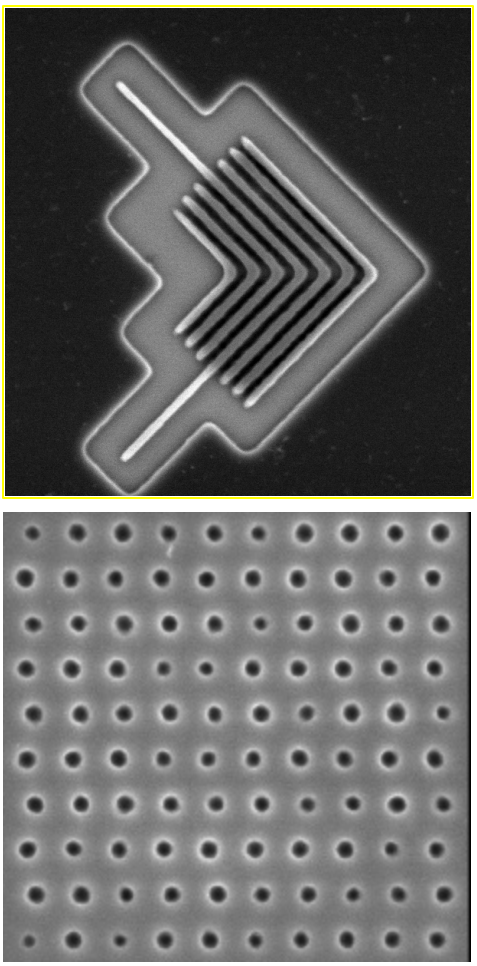


Figure 8: 70nm lines and contacts patterned with a 0.1NA prototype EUV exposure tool

below, there is significant concern as to whether EUV lithography will be ready for the 45nm generation. If EUV lithography is not ready, then 157nm lithography with Alternating Phase Shift Masks will be used for the 45nm generation.

In addition to 193nm, 157nm, and EUV lithography, Electron Projection Lithography (EPL) has been proposed for the 65nm node and below. There have also been proposals to use EPL as a complementary technology, specifically for patterning contact layers.

Although Intel continues to monitor the development of EPL technology, we do not see a place for EPL on the Intel roadmap. In particular, the low run rate of EPL tools will make the technology expensive. In addition, no one has demonstrated that full-size EPL masks can be fabricated with the low (zero) defect levels required for production. (There are similar concerns about defects on EUV masks. However, the mask industry has a clear, data-driven roadmap to achieve zero defect EUV masks in the 2005/2006 time frame when they will be required for process development). Finally, the successful patterning of 70nm contacts (Figure 8) with 0.10NA EUV optics show that a specialized tool for patterning contacts will not be required.

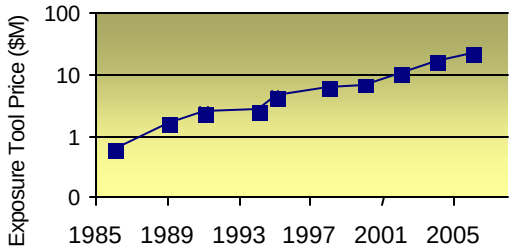


Figure 9: Exposure tool price trend

AFFORDABILITY

In 1986, Intel's first 150mm (6") factory was built and filled with manufacturing equipment for just over $25M. Today (2002) the typical price for a 193nm exposure tool is approximately $12M. The price of 157nm exposure tools is forecast to be as high as $20M; Extreme Ultra-Violet (EUV) exposure tools may cost as much as $25M (Figure 9). Fortunately, some of the price increases for lithography equipment have been offset by faster run rates (higher output per tool). As a result of higher tool output, the cost of Deep Ultra-Violet (DUV) lithography has actually decreased by ~20% since its introduction in the mid-1990s (Figure 10).

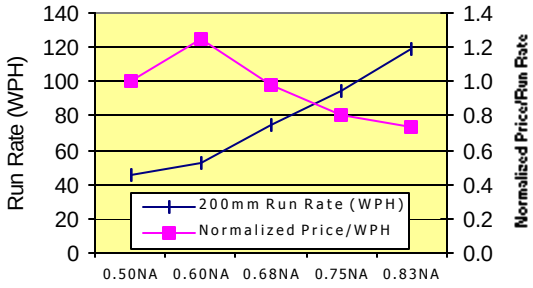


Figure 10: DUV exposure tool run rate trend

"Reuse" of lithography equipment allows the high cost of exposure tools to be spread over several generations of technology. Intel has a well-defined reuse "waterfall"

where tools that were originally purchased for patterning critical device layers are reused on subsequent process generations to pattern looser layers (Figure 11).

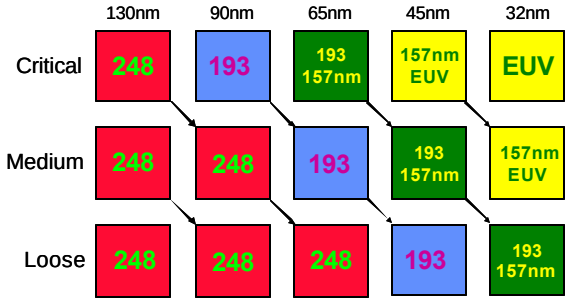


Figure 11: Intel reuse waterfall

Thus far, Intel has been able to maintain a fairly level cost for lithography by adopting the following strategy:

> Rapid transition to each new generation of lithography equipment; i.e., shorter wavelengths.
>
> Using fast (high-run rate) lithography tools.
>
> Reusing lithography equipment over multiple process generations.

Our expectation is that this strategy will allow lithography to continue to be affordable into the 45nm technology generation and beyond.

CONCLUSION

Although the transition to sub-wavelength patterning has accelerated the rate of introduction of new lithography technologies, the necessary technology does exist and will be available when needed by the semiconductor industry. In particular, 193nm lithography will be introduced into high-volume manufacturing in 2002. There are no technological barriers to the introduction of 157nm and Extreme Ultra-Violet (EUV) lithography in the 2005 to 2007 time frame. Finally, faster and higher-output exposure tools, combined with the practice of selective reuse of existing lithography equipment, will ensure that lithography remains affordable for the foreseeable future. There is no doubt that lithography will continue to play its pivotal role in enabling Moore's Law.

ACKNOWLEDGMENTS

The author acknowledges many valuable and heated discussions with the Intel SCS Lithography Core Team members, all of whom have made important contributions to the development of the Intel Lithography Roadmap.

EUV images were provided by the EUV LLC.

The 157nm images were used with the permission of the Willson Research Group at the University of Texas at Austin.

AUTHOR'S BIOGRAPHY

Peter Silverman is an Intel Fellow and Director of Lithography Capital Equipment Development. Peter joined Intel in 1978 and has held positions in process development, manufacturing, and engineering management. He is responsible for the coordination of Intel's Lithography Roadmap and for the technical and commercial management of lithography equipment development programs. Peter received a B.S. degree in Physics from MIT and a Ph.D. degree in Solid State Physics from the University of Maryland. His e-mail is Peter.J.Silverman@intel.com.

Copyright © Intel Corporation 2002. This publication was downloaded from http://developer.intel.com/.

Legal notices at http://www.intel.com/sites/corporate/tradmarx.htm.

Emerging Directions For Packaging Technologies

Ravi Mahajan, Technology and Manufacturing Group, Intel Corp.
Raj Nair, Technology and Manufacturing Group, Intel Corp.
Vijay Wakharkar, Technology and Manufacturing Group, Intel Corp.
Johanna Swan, Technology and Manufacturing Group, Intel Corp.
John Tang, Technology and Manufacturing Group, Intel Corp.
Gilroy Vandentop, Technology and Manufacturing Group, Intel Corp.

Index words: emerging directions, packaging, power delivery, vertical integration, thermal management, BBUL

ABSTRACT

The continual increasing performance of microelectronics products places a high demand on packaging technologies. Key drivers such as thermal management, power delivery, interconnect density, and integration require novel material development and new package architectures. In this paper, package technology migrations for microprocessors and communication products are described. Material needs for high thermal dissipation, high-speed signaling, and high-density interconnects are discussed.

Microprocessor scaling for increased performance and reduced cost places significant challenges on power delivery and power removal due to reducing dimensions, operating voltages, and increasing power. Meeting these challenges indicates a need for advanced packaging solutions, such as Bumpless Build-Up Layer Technology (BBUL); and power-delivery architectures such as On-Package Integrated Voltage Regulation (OPVR) that enhance the power-delivery capability of the packaging architecture. Similarly, solutions using advanced materials and heat management systems such as heat spreaders and high-capacity heat sinks are needed to facilitate power removal. Microprocessor scaling also requires improvements in package substrates and continues to drive major transitions in substrate materials and features while market constraints continue to exert significant cost pressures.

To support the ever-growing demand of cellular communication products for highly integrated, small form factor devices, new package architectures are described. Key research thrusts for the future are also highlighted.

INTRODUCTION

A review of the evolution of microprocessors in the past two decades and a projection into the near foreseeable future in the current decade shows that microprocessor performance continues to match the almost self-fulfilling prophecy of Moore's law [1]. This increase in performance places significant demands on packaging and assembly for performance and reliability. A paper published in the 3rd quarter of 2000 in the *Intel Technology Journal* [2] showed that in response to demand, microprocessor packaging has evolved from simple mechanical protection to a sophisticated electrical/thermal/mechanical platform that enables microprocessor performance. This paper elaborates further on the themes articulated in the earlier paper and provides additional details of some of the emerging trends in assembly and packaging. The key technical drivers for assembly and packaging in the areas of power delivery, power management, interconnect scaling, and integration are articulated. Future driver trends are discussed in order to explain some of the technical challenges these trends have created. Specific technical challenges in power delivery, thermal management, materials development, high-speed signaling, high-density interconnects, and integration are discussed, and the state of the art is reviewed. Opportunities for further work to continue to expand the cost-performance envelope of assembly and packaging technologies are identified; in particular the Bumpless Build-Up Layer (BBUL) technology is reviewed.

Attention is then shifted to the packaging of components used in portable and cellular devices. These applications demand low-cost, high-performance packaging in

compact form factors. The market segments present unique challenges in terms of cost, performance, and time to market. New package architectures developed to address these challenges are reviewed, and future developmental opportunities are highlighted.

POWER-DELIVERY CONSIDERATIONS

Microprocessor scaling has consistently adhered to Moore's law [1]. Increasing transistor density combined with the performance demanded from next-generation microprocessors result in increased processor power. Scaling also necessitates a reduction in the operating voltages both for reliability of the finer-dimension devices and for containing the power consumed. This reduction in the supply voltage further increases the supply currents drawn by the microprocessors while margins for noise in the power supply shrink in absolute terms.

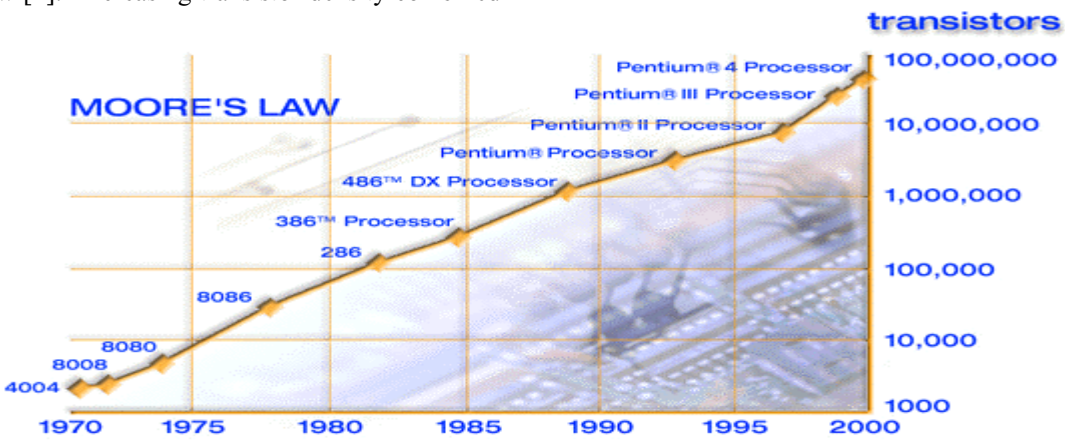


Figure 1: Intel CPU transistors double every ~18 months

The increasing supply currents and shrinking margins in supply noise place an enormous burden on the circuits that provide power to the chip. These circuits are collectively referred to as the power-delivery system for the processor, and they constitute the power-conversion devices or Voltage Regulator Modules (VRMs) that step a high voltage of 12 or 48V down to the processor operating voltage (~1.5V) as well as the hierarchy of capacitances located at the output of the VRM's extending all the way into the microprocessor package [3].

Increasing currents tax the power-delivery system of a CPU or network in two principal ways:

- Power-saving features in the CPU architecture mandate various operating conditions that lower power consumption to a minimum through 'sleep,' 'stand-by,' 'idle,' and 'power-down' states: when the CPU changes state to a fully operational mode, it demands a sudden surge in current in a short duration of time.
- The very large (>100A) currents flowing in the interconnect between the VRM's and the CPU cause power wastage and the associated self-heating in the board, socket, and CPU packages.

Power Path Loop Inductance Scaling and Reduction

The surges of current (often referred to in the literature as DI/DT events) as well as the sudden relaxation in the current demanded by the CPU, combined with the properties of the noise de-coupling capacitance hierarchy, result in a series of supply voltage variations referred to as supply droops and overshoots. The droops in the network are dependent upon the capacitance hierarchy, with the voltage variations being a consequence of damped resonant oscillations of the various de-coupling loops of capacitances and inductances in the network. It can be seen through simple analytical derivation that the magnitude of these droops can be calculated as follows:

$$\Delta V = \Delta I \sqrt{\frac{L_p}{C_d}} \qquad (1)$$

Where Lp and Cd in equation (1) refer to the package loop inductance and the die effective capacitance, respectively. The variation in supply voltage ΔV refers to the droop corresponding to the resonant response of the loop, including the die capacitance and the loop inductance, to the next level of de-coupling capacitance.

A key challenge in minimizing this droop is the scaling of the package loop inductance to meet the demands of CPU scaling. Exploring this further, we get the following:

$$\Delta V_1 = k_v V_1 = \Delta I_1 \sqrt{\frac{L_1}{C_1}} = k_i \frac{C_1 V_1^2 f_1}{V_1} \sqrt{\frac{L_1}{C_1}}$$

And

$$\Delta V_2 = k_v V_2 = \Delta I_2 \sqrt{\frac{L_2}{C_2}} = k_i \frac{C_2 V_2^2 f_2}{V_2} \sqrt{\frac{L_2}{C_2}}$$

Where X_i refers to a parameter X in the two processes. For example, L_1 is the package loop inductance in process-1 and L_2 is the same for the next-generation process-2. Defining S_l as (L_2 / L_1), and dividing the equations above, loop inductance scaling is given as

$$S_l = \frac{1}{S_c S_f^2} \qquad (2)$$

Where S_x refers to the scaling factor for parameter X. Interestingly, under the assumption that the die size and the architecture remain identical, the scaling factor for loop inductance is determined purely by the capacitance scaling factor (die capacitance per unit area) and the scaling factor for the operating frequency.

Power Progression for Intel CPUs

Historical data on the increase in power for Intel microprocessors is included in Figure 2. It is seen that the power doubles approximately every 36 months, which is approximately half the pace of the increase in the number of transistors as forecasted by Moore's Law. This difference could be attributed to greater amounts of memory content in microprocessors as they are scaled, leading to less overall capacitance contributing to power consumption than would be otherwise expected, or smaller die sizes. A capacitance scaling factor S_C = 1 reconciles the historical power trend with simple analytical predictions based on doubling frequency, transistor count, and less aggressive voltage reduction. These are described later.

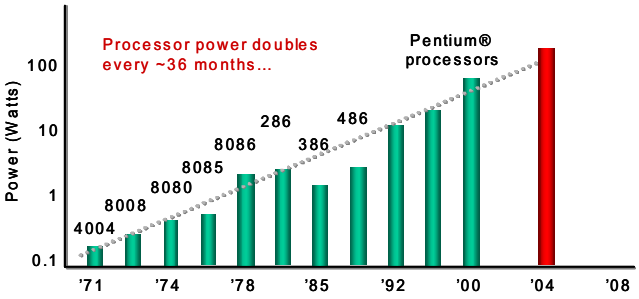


Figure 2: Historical power trend for Intel CPUs

Consider now a process scaling factor of $\frac{1}{\sqrt{2}}$ or 0.7 (as has been typical over a number of generations), and inspect power-related processor scaling scenarios:

- A *Quintuplet*[1] *'Q'* scenario, where C increases by (1/0.7) or S_c, the operating frequency increases by 2X or ~$(1/0.7)^4$, and the voltage reduces only by $\sqrt{0.7}$ or 85%, leading to a 2X increase in power and a $2^{\frac{5}{4}}$ or ~2.4X increase in the dynamic supply current.
- A *Realistic Scaling 'RS'* scenario where power P increases by a factor of $\sqrt{2}$ and the supply voltage V reduces by $\sqrt{\frac{1}{\sqrt{2}}}$ leading to an increase in the dynamic supply current of $2^{\frac{3}{4}}$ or ~1.68X.
- A *Triplet 'T'* scenario where the capacitance per unit area increases as before by S_c, but the frequency increases only by ~(1/0.7) or S_c, and the voltage reduces by a full 0.7 factor; the power in this scenario remains constant while the current increases by $2^{\frac{1}{2}}$ or ~1.43X.

It can be seen, by applying the simple scaling law (2) above that the *'Q'* scenario requires a scaling of the loop inductance by the 5th exponent of the process scaling factor S_c, while the *'T'* scenario also necessitates a scaling of the loop inductance by the 3rd exponent of the process scaling factor. The realistic *'RS'* scenario that reflects historical power scaling would require a loop-inductance scaling of the 4th exponent of process scaling.

[1] The motivation for the scaling scenario name will be evident in the consequence the scaling scenario imposes upon loop inductance scaling.

Figure 3: A Voltage Regulator Module (VRM)

As can be seen from equation (1), the knobs available to control the droops are limited from an assembly technology standpoint for the resonant loop of the highest frequency. Both ΔI and C_d are determined by the CPU architecture, circuit design, and layout as well as by process scaling. Components of assembly technology, particularly the package caps and the substrate, contribute to the effective loop inductance, L_p, that determines the droop in this damped resonant circuit. As supply currents increase and explicitly added die capacitances are removed (for die cost, leakage, and other die and circuit design reasons), reducing the power loop inductance to control droop ratios is a key driver that leads to the consideration of advanced packaging techniques such as the Bumpless Build-Up Layer (BBUL) technique, described in a later section.

Figure 4: Power pod power-delivery system

Power Path Series Resistance Scaling and Reduction

Power path resistance losses contribute significantly to system inefficiency as well as heat generation within the boards, sockets, and packages that support the microprocessor. The large current values anticipated in future microprocessors (in excess of 100A) are forcing greater pin counts in sockets, larger copper thicknesses in boards, and are arresting the thickness reduction (for aspect ratio control in fine trace widths) for metal layers within substrates. This is perhaps a more difficult challenge than the loop-L reduction.

Let's assume that the platform design requires that the loss in the power-delivery interconnect remain constant. As currents scale by a factor S_I, the resistance will need to be reduced by the square function of the current scaling factor to satisfy the constant interconnect power requirement, or

$$S_{ppr} = \frac{1}{S_I^2} \qquad (3)$$

where S_{ppr} is the power interconnect resistance scaling factor and S_I is the current scaling factor. It can be seen that the Q scenario for current scaling (where S_I = 2.4) will require a path resistance reduction by a factor of ~0.175 for the next process generation, the RS scenario requires a reduction factor of ~0.35, and the T scenario will need an S_{ppr} of ~0.5.

Figure 5 displays graphically the challenge in scaling assembly technology parameters in accordance with microprocessor scaling. Starting from hypothetical values for the loop-inductance and the series path resistance in the 0.18µm process generation, it is seen that the loop-L value required to meet the performance criteria three generations ahead is ~0.047pH for the Q scenario and 0.4pH for the T scenario. Similarly, the path resistance scaling beginning as before from a hypothetical value of 3.2 mΩ for the 0.18µm generation leads to a value of ~0.017 mΩ for the Q scenario and ~0.38 mΩ for the T scenario. The RS scenario lies somewhere in between these numbers. This illustrates the impracticality of scaling assembly technology parameters, as demanded by these scaling laws, in order that the performance requirements be met. It is becoming increasingly evident that increases in power or supply currents will require new architectures for power delivery to microprocessors in deep sub-micron processes.

While devices such as the VRMs shown in Figure 3 have advanced in their capability to deliver power, it is seen that solutions such as the Power-Pod (Figure 4), adopted for the anticipated high power and supply current numbers for the Intel Itanium™ family of processors will also be insufficient to meet the requirements of the 65nm node.

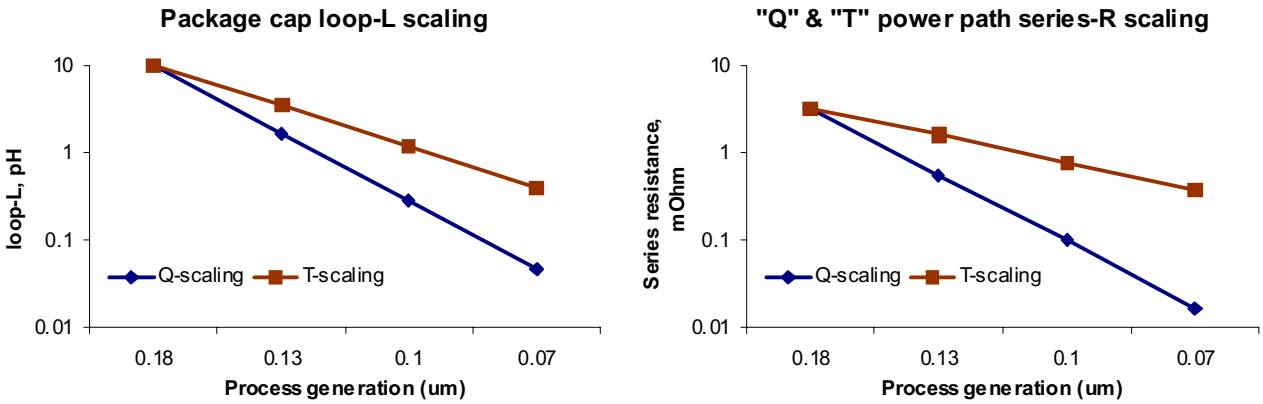


Figure 5: Scaling trends for package loop inductance and series path resistance

THERMAL MANAGEMENT CONSIDERATIONS

Technical challenges in the thermal management of microprocessors arise from two causes: (a) increasing power dissipation, which is concomitant with increasing performance; and (b) the need to cool regions of local power concentrations, often referred to as "hot spots." Typically, thermal management features are integrated in packages to spread heat while transporting heat from the die to the heat sink. The heat sink in turn dissipates heat to the local environment (see Figure 6 for a pictorial representation of this process).

The thermal management problem is one of ducting the Thermal Design Power (TDP) from the die surface at temperature Tj (commonly referred to as junction temperature) to the ambient at temperature Ta. In general terms, the temperature difference (Tj – Ta) is expected to slowly reduce over time since Tj can typically be forced lower by reliability and performance expectations, and Ta can be forced higher due to heating of the inside box air caused by increased integration and shrinking box sizes. Figure 2 shows TDP trends over time, and a simple scaling projection (as shown in the ITRS [4], for instance) indicates that the TDP will increase as a function of time, assuming no major design breakthroughs, which reduce microprocessor power, occur. Thus the thermal challenge arises from the fact that increasing values of TDP have to be ported between a diminishing temperature difference.

This challenge is exacerbated by another very important factor. On-die power distribution is typically not uniform. With increasing performance, the non-uniformity of on-die power distribution increases, and there are regions of the die dissipating high heat flux densities. These regions are commonly referred to as hot spots. Since the temperature of the hot spot can often affect performance and will always govern the overall reliability of the silicon, maintaining the hot spot temperature below certain limits is a requirement in thermal design. This leads to two undesirable consequences: (a) the focus on cooling the hot spot leads to a general over-design in the microprocessor cooling solution; and (b) the non-uniformity in the heat source limits the total amount of heat that can be managed by a thermal solution. The overall problem is graphically illustrated in Figure 7, which shows a plot of the overall cooling capability of different thermal solutions as a function of the Density Factor (DF), a factor defined in [5] as a measure of the impact of power non-uniformity.

Other significant constraints that must always be understood are the cost and integration constraints. While increasing power demands more sophisticated thermal management solutions, they have to stay within cost bounds dictated by the market segments in order to be economically viable. The solutions must also be capable of fitting within the chassis form factors and when assembled with the rest of the components, they should not reduce the reliability of the overall system.

Thermal Solution Strategies

An obvious solution is to mitigate the problem by design, i.e., design processors that are power efficient and have benign power topologies. This can be done by ensuring that thermal considerations are part of the design process rather than an afterthought.

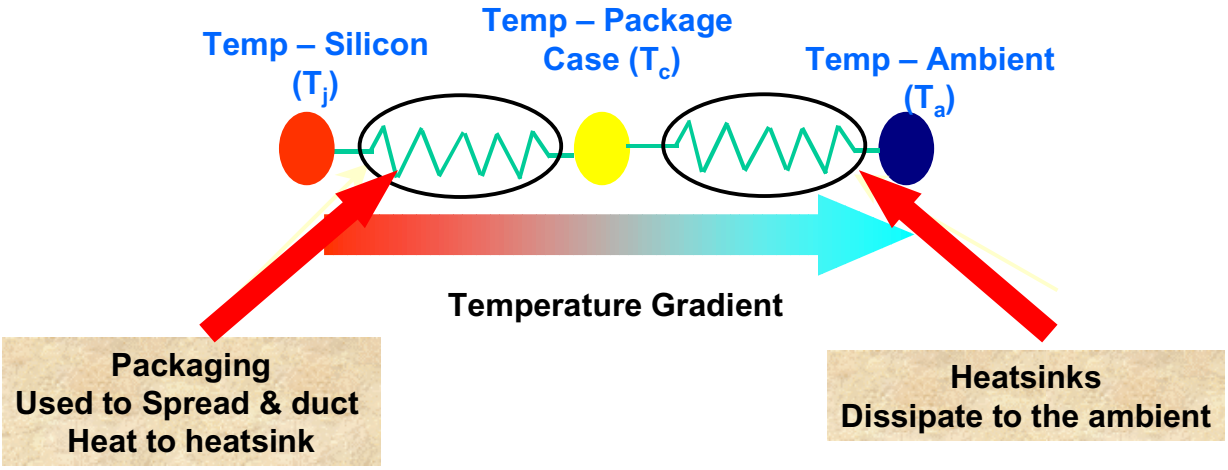


Figure 6: Temperatures and hierarchy in microprocessor cooling

Awareness of thermal issues and the need for thermal co-design has increased over the past few years as thermal management becomes one of the limiters of microprocessor performance. Thermally efficient designs could slow down the unconstrained increases predicted in Figure 2 and can help significantly in the development and deployment of cost-effective thermal solutions. However, while thermally benign designs are being developed, they cannot be depended on as the only strategy. Technology solution strategies need to be developed in parallel assuming that thermal demands will increase over time. These solution strategies can be broadly categorized as follows:

1. *Hot spot mitigation*: The goal of this would be to even out the temperature profiles due to non-uniform power distributions, as close to the source as possible, by spreading out the heat. The use of efficient and cost-effective methods to spread out the heat will increase cooling capabilities by reducing the Density Factor (DF) as seen in Figure 7. Spreading can be accomplished by optimal material and design development. One good example of this is shown in Figure 8. Figure 8 (a) shows an Integrated Heat Spreader (IHS) included in the packaging for Pentium® 4 processors, and Figure 8 (b) shows an integrated heat pipe lid developed for the Itanium™ processor.

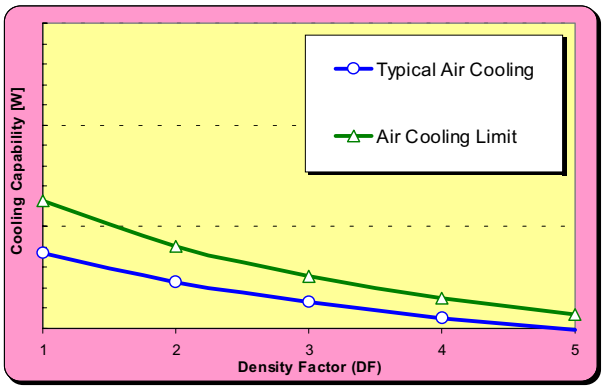


Figure 7: Impact of die power non-uniformity on cooling capability

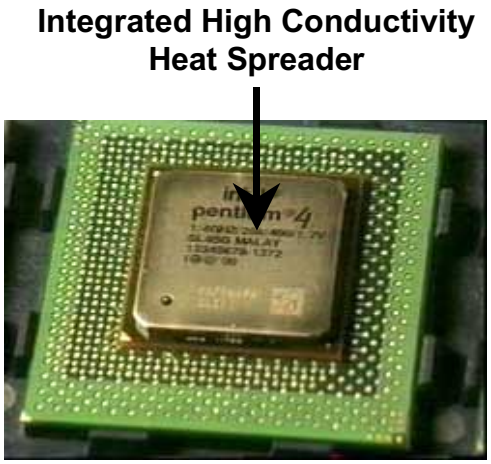


Figure 8(a): Use of IHS for Pentium® 4 processor

Pentium and Itanium are trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

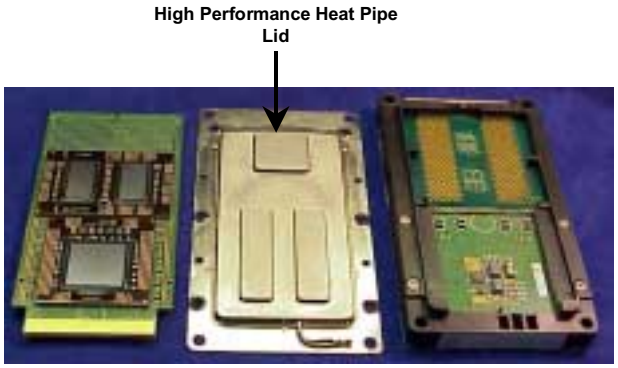


Figure 8(b): Heat pipe lid used for Itanium™ processor

2. *Increasing power-handling capability*: Currently air cooling techniques are used to cool microprocessors in most applications, i.e., a metal heat sink with air blowing over it is the primary cooling solution of choice. A representative analysis of the limits of air-cooling technology indicate that there are still some opportunities in air cooling that may be explored to increase cooling capability.

THERMAL MATERIALS TECHNOLOGIES

We can identify two basic architectures when describing heat removal for microprocessor packaging: (a) Architecture I (FC-XGA1), typically dealing with low-power (<30W) microprocessors or microprocessors in height-constrained applications, where the die is directly attached to the heat sink; and (b) Architecture II (FC-XGA2), typically dealing with medium- to high-power processors (>30W) where an Integrated Heat Spreader (HIS) is used to spread the heat. The term xGA applies to either PGA or BGA, and refers to the type of interconnect between the package and the motherboard. Figure 9 shows the basic implementation of these two architectures.

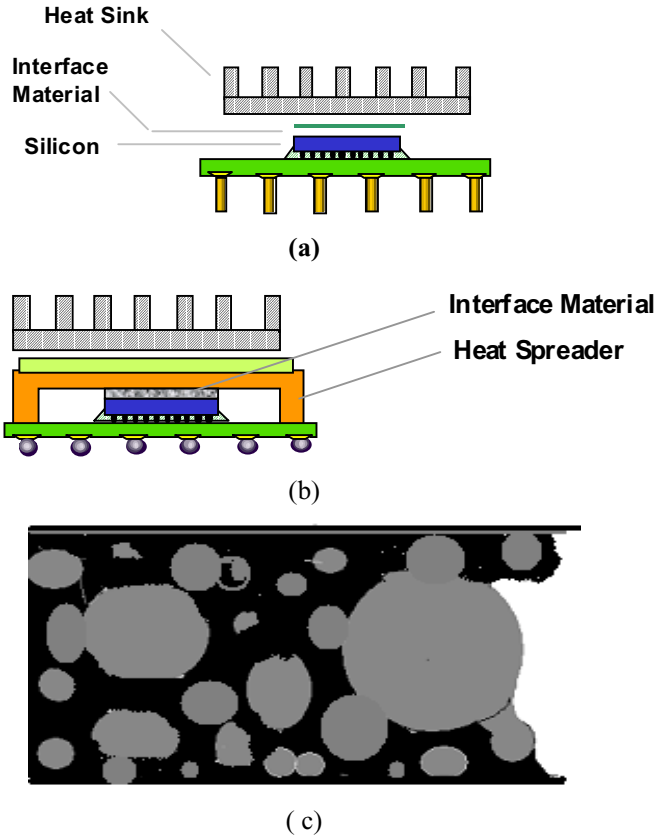


Figure 9: (a) Architecture I; (b) Architecture II; (c) schematic showing percolation in polymer TIM

In either case, successful thermal management requires the development of a Thermal Interface Material (TIM) that comes in contact with the die and heat sink (Figure 9 (a)) or the die and the heat spreader (Figure 9 (b)). Typically the TIMs are made up of a polymer matrix in combination with highly thermally conductive fillers (metal or ceramic) and can be classified as Phase Change Materials (PCM) and Thermal Greases and Gels [6]. Heat dissipation through these materials occurs through the phenomenon of percolation, schematically illustrated in Figure 9 (c). One can also consider providing thermal solutions for heat dissipation entirely through conduction, by utilizing TIMs made up of metals such as lead, tin, or bismuth. However, to integrate these metallic TIM material technologies with the silicon and packaging technology of choice poses an entirely new set of challenges from a stress, cost, and infrastructure perspective. An ideal case would be to develop a composite TIM material that provides enhanced heat dissipation while balancing the mechanical properties to minimize package stress. Developing these composite TIMs can possibly be achieved in a variety of ways including utilizing recent developments in nano-material technologies. It is

expected that one can develop materials with bulk thermal conductivities 5-10 times of those obtained using conventional approaches.

PACKAGE SUBSTRATE TECHNOLOGY

Package substrate technology has undergone significant changes in the past two decades. The early X86 processors were packaged in ceramic substrate with tungsten (W) or molybdenum (Mo) interconnects. Ceramic substrates continued to be the substrates of choice up to the Pentium® Pro microprocessor generation. However, ceramic substrates suffered from the disadvantage of high dielectric constants, thick dielectric layers (leading to thick packages), poor conductor materials relative to Cu (W or Mo), and limitations on feature size just to name a few. In the mid 1990s, Intel pioneered the transition from ceramic to organic substrates. Organic substrates provide better performance at a lower cost, and have evolved to be the substrates of choice for microprocessor packaging. Figure 10 illustrates the substrate evolution for Intel microprocessor generations.

As future microprocessor and network processors run at increased clock speeds, significant challenges are imposed on the performance of the substrate technology. The key technology drivers are as follows:

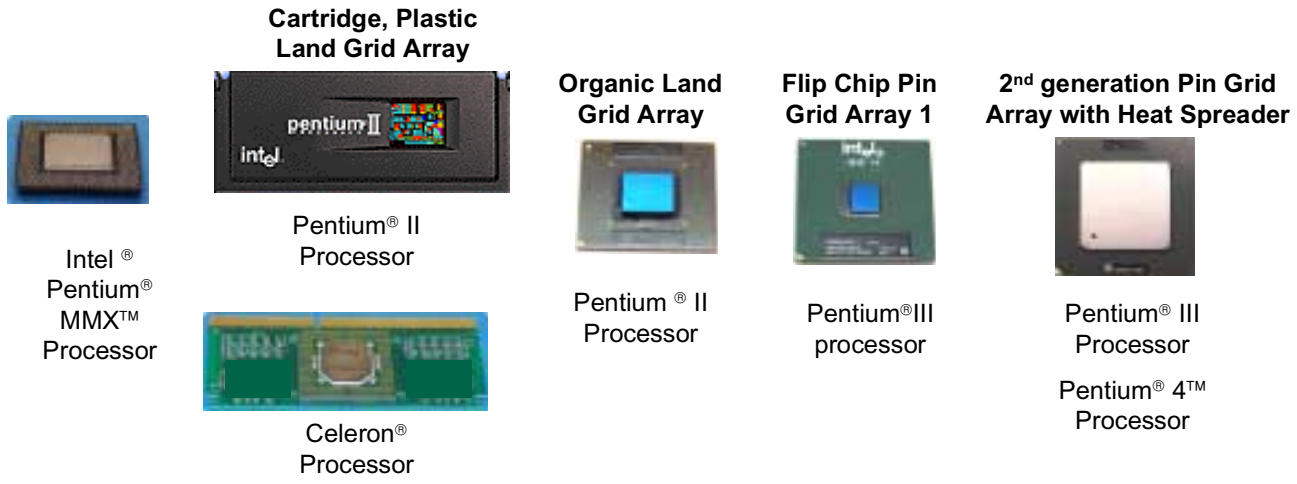


Figure 10: Substrate evolution

1) *Feature size reduction to increase routing density*. A reduction in feature size of the substrate includes a reduction in interconnect line width, line spacing, micro via diameter, capture pad diameter, plated through-hole diameter, flip-chip bump pad pitch, and flip-chip pad solder mask opening diameter. New processes and process materials are required. These include better photo resist and solder mask materials for higher resolution, better layer-to-layer alignment litho tooling and processes to reduce capture pad size, better mechanical drill bits to drill smaller plated through holes, better and higher through-put laser drill equipment for smaller micro vias, and better dielectric materials for improved metal adhesion. For today's state-of-the-art build-up processes, the line width and spacing is at about 25 microns. In about five years, the requirement will be substantially smaller.

2) *Increased performance*. In order to meet the requirements of future high-speed microprocessors, new dielectric materials and new process controls are needed. The new materials are expected to have low dielectric constants and low loss tangents. In addition, impedance matching and impedance control are critical for high-speed signals. Impedance control is a direct effect of dielectric thickness and line width tolerance; therefore, the control of line width and the tightening of tolerance are important parameters for future manufacturing processes. Improved processes and materials are also needed to reduce dielectric roughness while maintaining adhesion between the dielectric and copper lines in the substrate.

3) *Increase in mechanical loading*. Due to the increase in performance requirements, the power dissipation of future-generation microprocessors is rapidly increasing. As a result, larger and heavier heat sinks are employed to cool the microprocessor. In order to maintain intimate contact between the heat sink and package, a Thermal Interface Material

(TIM) is introduced between them. The two are then clamped so that the TIM is as thin as possible. This leads to fairly large static and dynamic loads on the package. In addition, the predicted eventual migration to Land Grid Array (LGA) sockets implies potentially higher static loading will be applied to the substrate. The substrate has to withstand all these mechanical loads through the life of the device. Another factor to consider is the implementation of Inner Layer Dielectric materials on the silicon with low dielectric constants (commonly referred to as low k ILD materials). These materials tend to become increasingly fragile with reduced dielectric constants. It is critical therefore to have a substrate material where the stress impact on silicon is mitigated while the thermo-mechanical reliability of the total system is maintained. This is another important requirement for the development of a new class of next-generation substrate materials.

4) *Thinner substrate*. The increase in demand for thin, portable, laptop computers drives the requirement for thinner substrates. This challenge can be addressed on two fronts. Firstly, optimize design with reduced feature sizes to reduce layer count. Secondly, develop thinner materials for the substrate core. Flex substrates using pliant polyimide materials are also being used to reduce thickness; however, cost is a key issue. Lower-cost flex materials are needed. Meeting the thin substrate requirement will require the industry to invest in new process equipment, handling equipment, and carriers. This will negatively impact the cost of the substrate.

5) *Lower Cost*. If price were no object, then it would not be as difficult to custom tailor a substrate technology to meet all the needs stated above. However, in reality, market pressures require that the future substrate costs must be reduced to ensure competitiveness in the marketplace.

In summary, the requirements of continued performance improvement, higher reliability, smaller features, and lower cost are driving the development of break-through technologies.

CHIP-TO-PACKAGE-LEVEL INTERCONNECT MATERIALS

Organic flip-chip technology is today's cost effective technology of choice for meeting high pin count and high-performance requirements. The ITRS roadmap [4] predicts the I/O pitch for the die-to-package interconnect to approach 120 microns in the next five years and 80 microns in the next ten years. A combination of decreasing pitch; environmental concerns (Pb/Halogen free); mechanical stress concerns (e.g., for low k dielectric integrity); electrical requirements (e.g., current density); and cost constraints are driving the development of bump and underfill materials technology in an entirely new direction. Environmental concerns will eventually lead to Pb-free flip-chip technology through the development of new bumping materials. A variety of material choices can be pursued for Pb-free bumping depending on the process of choice, e.g., plating, printing, stud-bumping etc. However, as highlighted in the ITRS, the bumping and underfill technology of choice will have to ensure low k dielectric integrity. A Pb-free bumping material with low yield stress and low creep resistance will be a critical enabler in the future.

The choice of the bumping material guides one towards the mechanical properties required in the underfill material (modulus, Coefficient of Thermal Expansion (CTE), etc.) to ensure adequate fatigue life for the bumps. Decreasing bump pitch and chip height and increasing bump density will eventually push the limits of capillary flow underfill materials. In order to achieve a breakthrough, one has to look for new directions in both polymer and filler technologies. Polymer resin technologies that can provide fundamentally low CTE (<35 ppm) and low viscosity, and that can be 'integrated' with the bumping materials of tomorrow will emerge as the resins of choice. Filler technologies that can be combined with this resin of choice to manage the underfill CTE without impacting flow and/or the bump-to-package substrate interconnect would be ideal. Underfill materials using such technologies can provide further opportunities to develop new cost-effective processes.

The ideal underfill material/process technology would have to be cost effective, and it would have to be capable of being scaled in such a way as to be independent of bump pitch and die size. This would point towards wafer-level underfills, the best case being where the underfill material/process can be integrated with the back-end Fab process technology. In summary, the flip-chip technology of tomorrow will drive a tighter coupling of silicon processing technology and assembly packaging technology. An alternate approach to meeting the needs of reducing pitch in the die package interconnect scaling is discussed in the next section.

BUMPLESS BUILD-UP LAYER PACKAGING

Bumpless Build-Up Layer (BBUL) packaging is a novel technology developed to meet future packaging technology requirements. It is constructed by fabricating the package layers on top of the chip as opposed to attaching a chip and package. The BBUL package provides the advantages of small electrical loop inductance and reduced thermo-mechanical strain imposed on interconnects with low k ILD materials. Furthermore, it allows for high lead count, ready integration of multiple electronic and optical components (such as logic, memory, radio frequency, microelectromechanical systems (MEMS), among others), and inherent scalability. A cross section of a 3-layer BBUL package is shown in Figure 11. Intel initiated the BBUL project with the goal of addressing scalability, power delivery, and low-k ILD compatibility issues. Detailed analysis on the advantages of the BBUL technology may be found in the published literature [7-9]. Here, we present a summary of the technical advantages and manufacturing challenges.

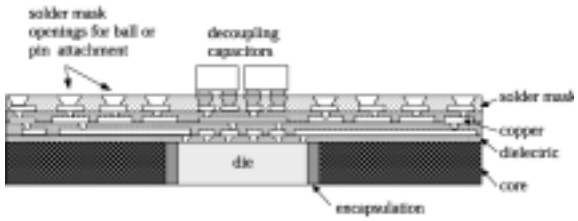


Figure 11: Schematic cross-section of a 3-layer BBUL package

The BBUL package shows a reduction in first droop, largely due to the decreased thickness of the package. BBUL package loop inductance is dominated by the inductance of the discrete decoupling capacitors; the package itself is a minor contributor to the loop inductance penalty. Another key advantage of BBUL packaging is in the area of mechanical reliability. The use of low-k ILD materials on the die is increasing the susceptibility of the die to mechanical failures caused by stresses imposed by the package. Figure 12 shows the relative out-of-plane stress for BBUL versus a standard package, as predicted using mechanical modeling [8] with the commercial finite element code ABAQUS (red is high stress, and blue is low stress on the rainbow scale) [10]. The figures show significantly reduced stress for the BBUL architecture. Equivalent comparisons for first principle stress and Von Mises stress are presented elsewhere [8]. We expect that the BBUL architecture will place lower stresses on the die, thus providing a mechanical advantage over a standard package.

BBUL also offers routability advantages over the standard package. Unlike many versions of flip-chip assembly (such as capillary underfill), die-package interconnections can be arbitrarily placed, as there are no restrictions imposed by an underfill process. This provides a significant advantage in the number of signals that can be routed out from the die on a single layer.

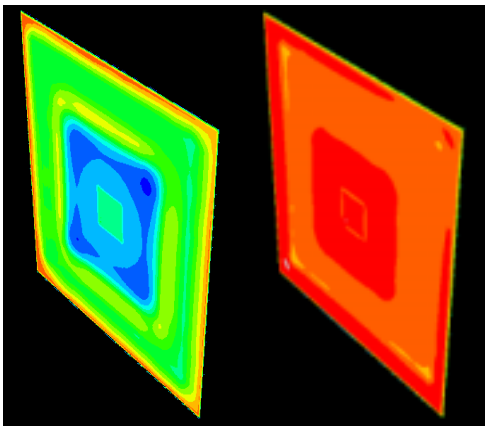


Figure 12: Out-of-plane die M6 to M7 stress for BBUL (left) and standard package (right)

In order to encapsulate the die inside the package, as is done in BBUL packaging, the process flow must deviate significantly from standard assembly. The process flow is shown in Figure 13. With standard assembly, the die and substrate are fabricated in parallel, tested independently, and finally assembled together to form the final package. With BBUL, substrate processing follows die processing, increasing the total throughput time and introducing known good die loss. These are significant manufacturing disadvantages and factor strongly in the final cost of the packaged die. Currently we are investigating, along with our package suppliers, methods of changing to the BBUL process flow and architecture in the hopes of retaining the performance advantages as shown above, while reducing the manufacturing penalty associated with the sequential process flow.

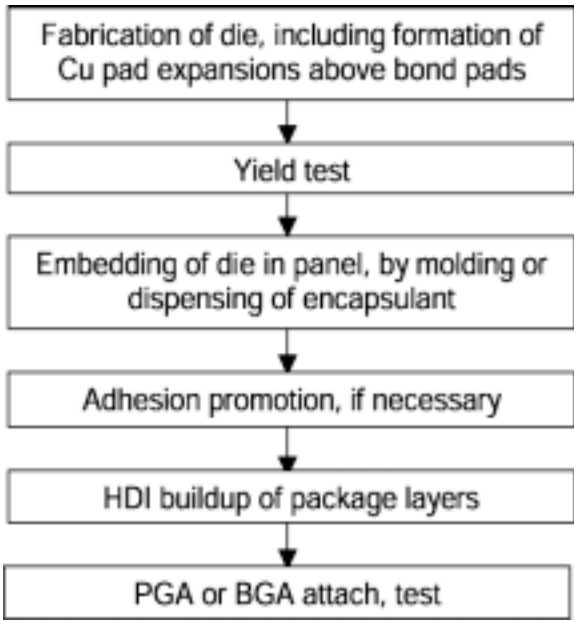


Figure 13: BBUL process flow

CELLULAR COMMUNICATION PRODUCTS

Another important direction for packaging is being driven by cellular communication products. Highly integrated, small form-factor packages are required to meet the demands of the emerging 3G application space. Integration in this environment goes beyond integration of flash with SRAM to integrating new memory (PSRAM, LPSDRAM) with a base band in addition to the traditional memory. In conjunction with the increase in the number of memory types, the available footprint on the board is shrinking, as seen in Figure 14. Vertical integration, i.e., the stacking of multiple die, is a typical approach to meeting the needs of this market segment [11]. While addressing in-plane constraints, vertical integration must also meet stringent height constraints. As a consequence of the increase in the types of die that are being stacked, an important package requirement is the flexibility to mix and change die within a package to meet demand and accommodate late changes. For example, the package architecture must allow for a memory upgrade without causing every layer of the package to be re-routed. These challenges are leading to new solutions beyond the current wire bond stacking found in present-day chip-scale packages.

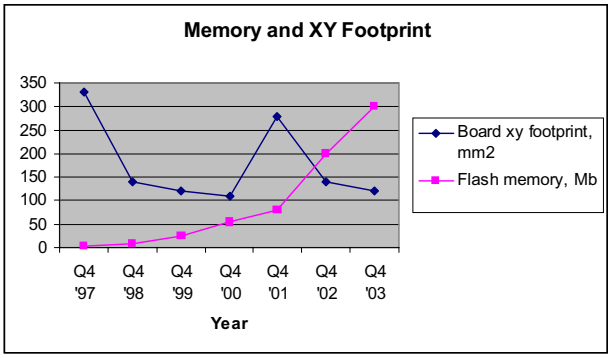


Figure 14: Memory increase and footprint on board

Current stacked-chip, scale-package architecture is shown in Figure 15. Technology efforts are underway to minimize the thickness of each component within the package, and the trend in substrate and die thickness is shown in Figure 16.

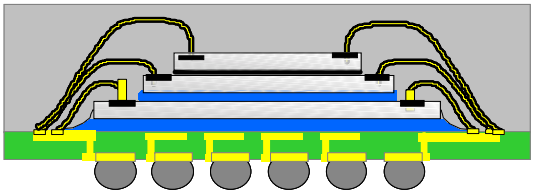


Figure 15: Current stacked CSP (1.4mm)

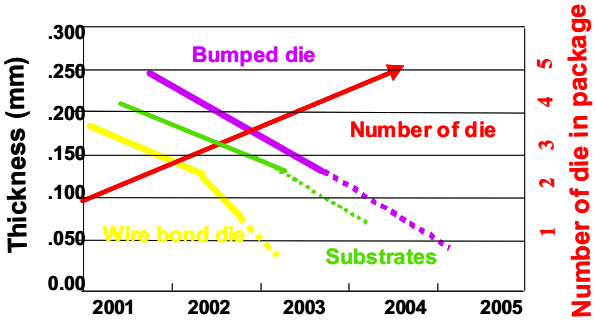


Figure 16: Trends in die/substrate thicknesses

As limits are approached on minimizing the die thickness, forming an interconnect through the silicon is another key direction [12]. The through silicon interconnect allows for a minimum in-plane and vertical footprint of the package. The vias are formed as part of the backend wafer processing, and typical dimensions are on the order of 10-50 microns in diameter with die thicknesses ranging from 25-150 microns.

There are two key issues that drive an alternate package architecture away from die-to-die stacking and towards package-to-package stacking. The first issue is the configurability of the different die within a stack. The

number of possible die combinations is growing exponentially. At the same time the need to respond to changes from the customer do not allow for long re-designs to upgrade memory or stack the different memory with the same base band. A stack-stack architecture can be designed that allows for minimizing the required changes by standardizing on-pin locations in the package-to-package interconnect. The top layer of the package can be quickly reconfigured without impacting the other two packages in the stack, as long as the interconnect terminals supply the same functionality. The second key issue is stacking a mix of Known Good Die (KGD) with non-KGD and the associated yield loss. Sensitivity to yield losses increases with the number of die in a stack. Memory can typically be completely tested by adding a third Sort step, while a base band cannot be as economically tested at the wafer level. Therefore, the ability to package and test the die prior to stacking allows for minimizing die yield losses. Assembly yield also may dictate the need to perform an open/short testing as the packages are stacked. The mix and number of die within a stack will eventually dictate the decision to use die/die stacking versus package/package stacking architectures.

SUMMARY

Packaging is one of the key enablers for microprocessor performance. As performance increases, the technical challenges in the areas of power delivery, interconnect scaling, interconnect performance, power removal, and mechanical reliability increase. This in turn requires the development of new materials and package architectures to enable microprocessor performance. A general overview of the emerging trends has been presented in this paper. An attempt has also been made to provide a context for some of the cost and integration constraints imposed by market conditions on the choice of materials, packaging architectures, and form factors.

Similarly, some of the unique requirements in the cellular market segment due to various form factor and integration requirements have been discussed to provide insight into some of the emergent packaging trends.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the encouragement and guidance in this work received from Bala Natarajan and also thank Mostafa Aghazadeh, Gerald Marcyk, and Nasser Grayeli for reviewing this document. Finally, we would like to thank the technical experts in the ATD Design, ATD Materials and Component Research groups for some of the analyses used to identify the emerging directions.

REFERENCES

[1] Moore, Gordon E., "Cramming More Components into Integrated Circuits," *Electronics*, Vol. 38, No. 8, April 1965.

[2] Ravi Mahajan, Ken Brown, and Vasu Atluri, "The Evolution of Microprocessor Packaging," *Intel Journal of Technology*, 3rd Quarter, 2000. http://developer.intel.com/technology/itj/q32000.htm

[3] Nair, Raj, "Assembly Technology Pathfinding Challenges," *Intel Assembly & Test Technology Journal*, 2001.

[4] ITRS 2001 Roadmap, www.sematech.org

[5] Chia-pin Chiu, Javier Torresola, Greg Chrysler, Dean Grannes, Ravi Mahajan, and Ravi Prasher, "Density Factor Approach to Representing Impact of Die Power Maps on Thermal Management" in preparation.

[6] Ram Viswanath, Vijay Wakharkar, Abhay Watwe, Vassou LeBonheur, "Thermal Performance from Silicon to Systems," *Intel Journal of Technology*, 3rd Quarter, 2000, http://developer.intel.com/technology/itj/q32000.htm

[7] Steve Towle, Henning Braunisch, Chuan Hu, Richard Emery, and Gilroy Vandentop, "Bumpless Build-up Layer Packaging," in *Proceedings ASME Int. Mech. Eng. Congress and Exposition* (IMECE), New York, Nov. 11-16, 2001, EPP-24703.

[8] R. Emery, S. Towle, H. Braunisch, C. Hu, G. Raiser, and G. J. Vandentop, "Novel microelectronic packaging method for reduced thermo mechanical stresses on low dielectric constant materials," in *Proceedings Adv. Metallization Conf.* (AMC), Montreal, Oct. 9-11, 2001, 7 pages, in press.

[9] H. Braunisch, S. Towle, R. Emery, C. Hu, and G. J. Vandentop, "Electrical performance of bumpless build-up layer packaging," in *Proceedings IEEE Electronic Components and Tech. Conference*, (ECTC), San Diego, May 28-31, 2002, in press.

[10] Hibbitt, Karlsson, and Sorensen, *ABAQUS* Version 6.2 User's Manual, Vol. 2, pp. 13.1.2-2, Pawtucket, RI.

[11] Smith, Lee and Zoba, David, "An Extremely-Thin Profile, Ball Grid Array Style Chip Scale Package," in *Proceedings of the SMTA International Conference*, Chicago, Illinois, September 2000.

[12] Yoshihiro Tomita, Tadahiro Morifuji, Tatsuya Ando, Masamoto Tago, Ryoichi Kajiwara, Yoshihiko Nemoto, Tomonori Fujii, Yoshifumi

Kitayama, and Kenji Takahashi, "Advanced Packaging Technologies on 3D Stacked LSI Utilizing the Micro Interconnections and the Layered Microthin Encapsulation," in *Proceedings of ECTC* 2001.

AUTHORS' BIOGRAPHIES

Ravi Mahajan received his B.S. degree from the University of Bombay in 1985, his M.S. degree from the University of Houston in 1987, and his Ph.D. degree from Lehigh University in 1992, all in Mechanical Engineering. He is currently in ATD Pathfinding and is primarily responsible for working with different groups within Intel Corp. to set strategic directions for thermal management of microprocessors. He is also a program manager responsible for establishing assembly and packaging strategies for the next-generation microprocessors. In addition, he is the Intel representative on the Technical Advisory Board for the Packaging and Interconnect thrust within the Semi-conductor Research Corporation. Ravi joined Intel in 1992 as a stress analyst in ATD. He then moved on to manage an analysis group and a TM Lab that developed a series of sophisticated analytical and experimental tools for design development and reliability assessments. He holds nine patents and is the author of several technical papers. His e-mail is ravi.v.mahajan@intel.com.

Raj Nair obtained his B.E. degree from the University of Mysore, India in 1986 and his M.S.E.E. degree from Louisiana State University in 1994. He joined Intel in 1995. He was the architect, designer, and implementer of Intel's first on-chip distributed regulation system. He then went on to be the architect and designer of an image sensor chip, and he researched CPU clocking, power delivery, packaging and signaling at Intel's microprocessor research labs before joining ATD-Pathfinding. He is currently responsible for project management in strategic initiatives enabling processor power delivery and IO. Prior to joining Intel, Raj spent about eight years developing and implementing machine automation systems, test and measurement instrumentation, and signal conditioning and data acquisition systems. Raj holds nine US patents and has numerous publications pertaining to voltage regulation, digital image sensing, digital clock distribution, and high-speed signaling. His e-mail is raj.nair@intel.com.

Vijay Wakharkar joined Intel Corporation in January 1991. Vijay received his B.S. degree in Metallurgy from The College of Engineering, Poona, India in 1982 and his Ph.D. degree in Materials Science and Engineering from SUNY at StonyBrook in 1989. He is currently managing the Materials group responsible for polymers and heat spreader materials and supplier development within the Assembly Technology Development Group. Vijay has worked at Intel for eleven years on materials development projects supporting the various package technology efforts within ATD ranging from TCP, PPGA, PLGA, Cartridge (SECc), and Flip-Chip Technology. Prior to working at Intel, Vijay spent two years as a Post Doctoral Associate at the IBM Alamden Research Center in San Jose. His e-mail is vijay.s.wakharkar@intel.com.

Johanna Swan received her B.S. degree in Mechanical Engineering in 1984 from Northern Arizona University. She joined Intel in 2000 as part of Assembly Technology Pathfinding. She works in the areas of stacked die and stacked packaging, as well as in the area of optical packaging to find solutions for emerging technology needs in the wireless communications arena. Before joining Intel she worked at Lawrence Livermore National Lab for sixteen years on projects ranging from X-ray diagnostics, laser beam delivery, magnetics and EUV lithography. Her e-mail is johanna.m.swan@intel.com.

John Tang received his B.S. degree in Chemical Engineering from the State University of New York at Buffalo in 1981 and his M.S. degree in Chemical Engineering from Northwestern University in 1983. He joined Intel in 1995 in the Assembly Test Materials Organization. He worked with substrate suppliers on the technology development and certification of the PLGA substrate. He also worked on the HVM ramp of PLGA in Intel factories worldwide. He then joined the Assembly Test Subcontract Group as a project manager, and was responsible for managing the assembly subcontractors for Intel's non-CPU products. He is now working at Assembly Technology Development pathfinding on substrate pathfinding and a long-range roadmap. Before Intel, John worked for IBM for eleven years. His experience included PCB board manufacturing, PCB board development, SMT development, Main Frame assembly development, direct chip attach development, and supplier management. His e-mail is john.tang@intel.com.

Gilroy Vandentop received his B.Sc. degree in Chemistry from the University of Alberta in 1986 and his Ph.D. degree in Physical Chemistry from U.C. Berkeley in 1990. He currently manages a packaging research group in Chandler, AZ, within Intel's Components Research organization. His group works in the areas of optical packaging and other novel packaging architectures to enable advancements in thermal, mechanical, and electrical package

performance. Before joining Components Research in packaging, Gilroy worked in Portland Technology Development for ten years, working on silicon process development in the areas of etch and photolithography. His e-mail is gilroy.vandentop@intel.com.

Copyright © Intel Corporation 2002. This publication was downloaded from http://developer.intel.com/.

Legal notices at
http://www.intel.com/sites/corporate/tradmarx.htm

For further information visit:

developer.intel.com/technology/itj/index.htm

Copyright © 2002, Intel Corporation. All rights reserved.
Intel is a trademark or registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.
For a complete listing of trademark information visit: www.intel.com/sites/corporate/tradmarx.htm